AN IMPROVED THUNDERSTORM FORECAST
INDEX FOR CAPE CANAVERAL, FLORIDA

THESIS

James A. Everitt, Captain, USAF
AFIT/GM/ENP/99M-06

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AN IMPROVED THUNDERSTORM FORECAST INDEX

FOR CAPE CANAVERAL, FLORIDA

THESIS

James A. Everitt, Captain, USAF

AFIT/GM/ENP/99M-06

DTIC QUALITY INSPECTED 2

AFIT/GM/ENP/99M-06

AN IMPROVED THUNDERSTORM FORECAST INDEX FOR

CAPE CANAVERAL, FLORIDA

THESIS

Presented to the Faculty of the Graduate School of Engineering

of the Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Meteorology
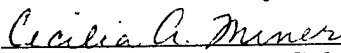
James A. Everitt, B.S.

Captain, USAF

March 1999

AN IMPROVED THUNDERSTORM FORECAST INDEX FOR

CAPE CANAVERAL, FLORIDA

James A. Everitt, B.S.

Captain, USAF
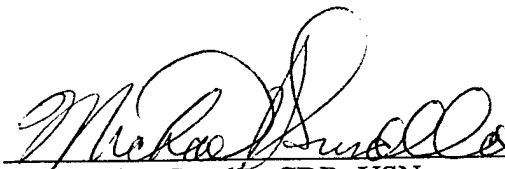
Approved:

Cecilia A. Miner   *4 Mar 99*
_____   _____
Cecilia A. Miner, Lt Col, USAF   Date
Chairman, Advisory Committee

Michael K. Walters   *4 Mar 99*
_____   _____
Michael K. Walters, Lt Col,, USAF   Date
Member, Advisory Committee

Michael P. Susalla   *4 Mar 99*
_____   _____
Michael P. Susalla, CDR, USN   Date
Member, Advisory Committee

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

This thesis creates a new algorithm to replace the Neumann-Pfeffer Thunderstorm Index (NPTI). The NPTI was created to provide an objective means of determining the probability of a thunderstorm occurrence. The 45th Weather Squadron at Patrick AFB, Florida, uses the NPTI to provide the probability for the occurrence of thunderstorms at Cape Canaveral. The probability is used for mission planning and resource protection, and increasing the accuracy of NPTI can potentially save billions of dollars for the United States space program.

Stratified logistical regressions are performed and probability equations are derived for May through September using upper air data and surface observations for Cape Canaveral. A logistical regression of NPTI was also performed. Variables include combinations of the climatological frequency of thunderstorms, the conditional probability of thunderstorms, the u and v components of the 850-mb, 700-mb, and 600-mb winds, the 800-mb to 600-mb mean relative humidity, the K index, and the Thompson index. The two resulting algorithms are compared to NPTI and persistence, and are evaluated based on their ability to forecast thunderstorms correctly. The primary performance metrics used to evaluate the algorithms are hit rate, threat score, probability of detection, false alarm rate, Brier skill score, and ratio skill score.

The investigation results indicate that the new algorithms are suitable for use by the 45th Weather Squadron and are an improvement on NPTI. The results show that the best algorithm, Stratified Logistic Thunderstorm Index (SLTI), has a 57% better hit rate,

a 51% better threat score, and a 68% better probability of detection than NPTI. In addition SLTI shows a 59% lower false alarm rate than NPTI. Because of the significant improvement over NPTI, the algorithm should be prepared for operational use at Patrick AFB.

1. *Introduction*

Weather has always played an important part in launch and recovery operations in the aerospace field, but in recent years the accuracy of forecasts has become paramount. Accidents resulting from launches during poor weather have highlighted the need for good forecasts. Fundamentally, accurate predictions are needed to maximize safety and reduce costs. Therefore, a review of current techniques with the intent of improving accuracy can have a real value, especially if an improvement is realized. This research project focuses on one method used in the prediction of thunderstorms and has the goal of increasing its accuracy.

1.1 *Overview*

Cape Canaveral, Florida, is the origin of many of this nation's space missions and these missions are very sensitive to weather conditions. Therefore, weather thresholds play a large part in daily operations. See Figure 1 for the geographic location of Cape Canaveral. In addition to the forecasts for normal resource protection, each launch vehicle has its own forecast criteria. One criterion of special interest is the probability of having a thunderstorm on station. Typically the 45th Weather Squadron (WS) at Patrick Air Force Base is responsible for providing the daily thunderstorm forecast. The 45th WS uses the Neumann-Pfeffer Thunderstorm Index (NPTI) to predict the possibility of exceeding the thunderstorm threshold. The NPTI is an algorithm which, when supplied with five input variables, calculates the probability of a thunderstorm being reported on station on the current day. The NPTI is designed to be valid for the current day, and decisions about operations are made from this probability.

1

Figure 1 Map of Florida

The NPTI was created over 25 years ago and recent updates brought to light some of its inaccuracies (Howell, 1998). This thesis focuses on finding a more accurate algorithm for calculating the probability of a thunderstorm.

1.2   Background

Weather affects all aspects of space operations, including rolling out, launching, and recovering vehicles. Unfortunately, the Cape Canaveral complex is located in an area which generates one of the highest frequencies of thunderstorms in the world (Falls et al., 1971). This means a high priority must be placed on thunderstorm forecasting.

1.2.1   Thunderstorms

The high frequency of thunderstorms in this area is related to the presence of all the necessary ingredients for thunderstorm formation: moisture, instability, and lift. The first of these, moisture, is abundant throughout Florida. This enhanced moisture is due in part to the presence of two major bodies of water on either side of the state. However, Cape Canaveral has additional moisture available because of its location on a peninsula surrounded by rivers. Figure 2 shows the extent of the availability of water. The Cape's southerly location also affects the second ingredient, instability. The warmer temperatures
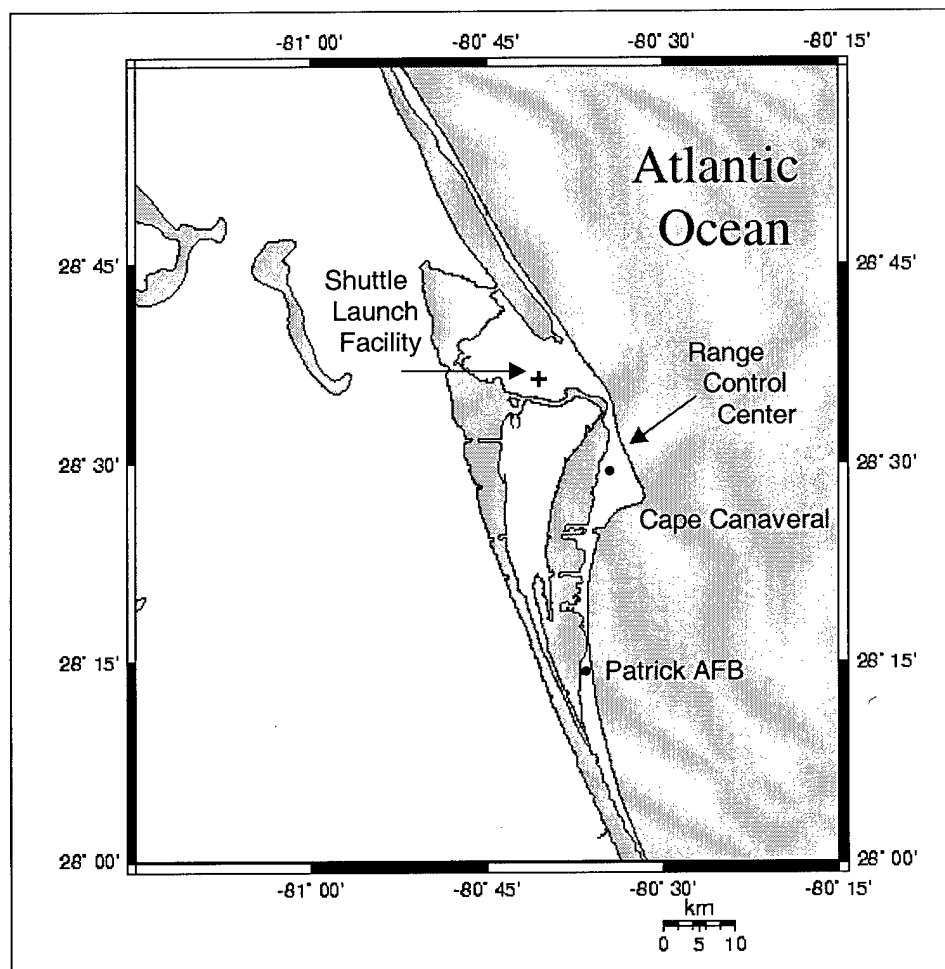
Figure 2 Map of Cape Canaveral

3

of the subtropics supplement the warm summer temperatures and increase the instability of the atmosphere. The third ingredient, lift, is provided by more complex mechanisms. Easily tracked and analyzed synoptic features normally provide lift, but during summers in Florida, these features are rare (Byers and Rodebush, 1948). When there are no synoptic features, smaller scale interactions--primarily the sea breeze--take over. How the sea breeze interacts with the local environment becomes a predominant parameter in determining lift and therefore predicting thunderstorms (Byers and Rodebush, 1948).

1.2.2   *Neumann-Pfeffer Thunderstorm Index (NPTI)*

Forecasters at the 45th WS, at Patrick Air Force Base, use the Neumann-Pfeffer Thunderstorm Index (NPTI) as the primary tool to determine the probability of thunderstorms. The NPTI algorithm, developed in 1971 by C. J. Neumann, calculates the probability of a thunderstorm and outputs a yes or no thunderstorm forecast. It uses five input variables from the morning (1200 Universal Coordinate Time (UTC)) radiosonde. They are the daily climatological thunderstorm probability, the 850-mb winds, the 500-mb winds, the 600-mb to 800-mb mean relative humidity, and the Showalter Stability Index. Each variable has been regressed linearly by month against the dependent variable, daily thunderstorm frequency. This frequency is the day-by-day record of whether or not a thunderstorm occurred. The coefficients obtained from the regression of each variable are used as variables in a further regression (Neumann, 1971; Hope and Neumann, 1972). The NPTI was developed using data from 1951 through 1969 in a statistical analysis. In 1987, a correction was applied to the coefficients in the algorithm, and the form with the corrected coefficients is the one used currently (Roeder, 1998).

4

1.3 Statement of the Problem

How can forecast accuracy be increased? The inclusion of an additional two years of data into the regression analysis has been shown to offer no significant improvement in the accuracy of NPTI (Howell, 1998). Furthermore, it has been shown that the NPTI has no increase in accuracy over forecasting with persistence (Neumann, 1971; Howell, 1998). Persistence uses the last occurring event as the forecast for the next event. This thesis reports on three alternative methods to improve NPTI's forecast accuracy. The first method uses alternative variables in the regression to obtain a new algorithm. The second method uses nonlinear regression techniques. Finally, the third method separates the data into categories and regresses the categories separately.

1.3.1 Objectives

The goal of this thesis is to improve the objective forecast accuracy of afternoon thunderstorm predictions at Cape Canaveral. The aim is to create an algorithm with improved accuracy over current methods of forecasting thunderstorms. Because persistence performs better than NPTI, an improvement over persistence corresponds to an improvement over NPTI. Therefore, a hit rate of 76% is necessary to be 95% confident of improving on the 71% hit rate of persistence found in this study. The increased accuracy would lead to increased safety of operations and a reduction of costs incurred by avoiding decisions based on incorrect forecasts.

1.3.2 Scope

This thesis is limited to the study of predicting summer thunderstorms at the Cape Canaveral complex in eastern Florida. Thunderstorm occurrence, from 1950 to 1996, recorded by the official observation site and the lightning detection system are used for the dependent variable in a regression. While the observation site has moved several times, these small geographical shifts are considered insignificant (Neumann, 1970). The research includes only thunderstorms that occurred during the convective season, which for the purposes of this thesis, is defined as the summer months from May through September. This period also corresponds to the period for which Neumann created the NPTI, and is also comparable to timeframes of Howell's study (Howell, 1998). During this convective season, there is a shortage of transient synoptic mechanisms that cause lift. This lack of synoptic forcing allows the probability of the thunderstorms to be closely linked to the state of the local environment as opposed to the state of an air mass that moves into the region.

The data for the study includes surface observations and radiosonde data from 1950 to 1996. Radiosonde data from 0900UTC to 1600UTC represent the upper air environment. Surface observations from 1100UTC to 0400UTC (0700-2400L) are used to determine thunderstorm frequency and for verification of regression algorithms. Ninety percent of the data is used to perform regression and analysis while ten percent is for verification. The verification data is randomly selected and removed from each database. The final algorithm predicts the probability of a thunderstorm at Cape Canaveral for a given day. This forecast is valid from 1100UTC to 0400UTC (0700-2400L).

1.3.3   Benefit of solving the Problem

Every year, various mission activities at Cape Canaveral are delayed or cancelled due to weather. Correct weather forecasts can reduce the number of unnecessary delays and cancellations, while incorrect forecasts can increase the physical and capital risk involved with operation in bad weather. In the Cape Canaveral area, there are over $8 billion in facilities. Furthermore, there are over 5000 pre-launch operations per year; and, of an average 60 attempted launches per year, only 35 are successful (Roeder, 1998). Also, the exposed nature of much of the outdoor work necessary for launch operations makes weather an important decision factor when considering the safety of personnel. It is estimated to cost $1 million just to de-fuel then prepare the space shuttle again after a thunderstorm is forecast (Roeder, 1998). In addition, because of the nature of how launch vehicles acheive orbit, missing a launch window can mean missing an opportunity, possibly forever. Therefore, reducing the number of delays or cancellations due to incorrect forecasts is a high priority. Increased accuracy can save lives, money, and time.

1.4   Procedure

The research for this thesis involves three main tasks. These tasks include data analysis and manipulation, regression, and verification. The first task, data analysis and manipulation, is an important part of preparing for the regression of the data. Without good preparation of the data, relationships between variables could be overlooked. The second task uses the data to obtain an algorithm for forecasting thunderstorms, and the

third task evaluates the algorithms created. S-plus® and Mathcad® are used for all statistical calculations.

The first task can be broken down into organizing the dependent data, searching for independent variables, finding correlations between variables and the dependent data, and finding categories in which to regress the data. Once the data are obtained from the observations, the measured values are used as variables, or further variables are derived from combinations of them. Unfortunately, some measured values are missing from the observations. These missing values are replaced by interpolated values. Next, the correlation between variables and thunderstorm frequency is determined. This correlation provides insight into which variables work best for regression. Finally, the data is divided into various categories. Each category, such as wind sector or month, is regressed separately. This stratification allows the accuracy of the regression to be maximized for each division.

The second task, regression, provides the algorithm for use by the 45th WS. Accuracy is improved by changing the method of regression. Neumann used linear regression for a linear fit, but a curvilinear fit is better because the dependent variable is dichotomous. Every value in the set of dependent variables is either 1 or 0. Because of this dichotomous nature, data is regressed against a logistic distribution which has a maximum of 1 and a minimum of 0. To compare the forecast ability of an algorithm created with logistic regression to an algorithm created with linear regression, the logistically regressed algorithm, Logistic Neumann-Pfeffer Thunderstorm Index (LNPTI), is computed using the same variables as NPTI. Another algorithm, the

Stratified Logistic Thunderstorm Index (SLTI), is created by using the regressions derived from the stratified data.

The final task, algorithm verification, determines which method is recommended for operational use. Verification is accomplished using the contingency table method (Wilks, 1995). The forecasts from each algorithm are verified with the independent data set and compared with persistence and the current NPTI. Only an algorithm showing significant improvement in forecast accuracy over the current NPTI is acceptable. The algorithm with the best skill score is the one recommended for use.

1.5   *Summary of Results*

Three new algorithms were created to forecast afternoon thunderstorms. They were objectively verified using standard measures of accuracy and skill scores. Four hundred thirteen independent events were used as a verification data set. The verification data was different than the regression data used to create the algorithms. The algorithm found to be the best, the Stratified Logistic Thunderstorm Index (SLTI), achieved a 78% hit rate. Using Ratio Skill Scores, STLI shows a 49% improvement over NPTI and a 24% improvement over persistence. All other measures of accuracy showed similar results. The results show that the new algorithms produce forecasts that are correct more often than either NPTI or persistence.

2. *Literature Review*

Studying thunderstorms is an ongoing project in the field of meteorology. This chapter will review previous research and explain some of the theoretical concepts used to forecast thunderstorms. The topics addressed include the generation of thunderstorms in Florida, current thunderstorm prediction techniques, and statistical theory used in this project.

2.1 *Thunderstorm Genesis*

A large number of studies have been made of the thunderstorm activity on the Florida peninsula. Because of the large number of studies, almost every meteorological parameter has been recorded, tracked, analyzed, or graphed at some point. These studies have provided a much greater understanding of the processes that cause thunderstorms for this area. Some increase in forecasting accuracy can be attributed to improved observational capabilities such as better radar and satellite coverage. However, part of the increased forecasting accuracy comes from a better understanding of the meteorological processes involved. This section reviews some of the findings of these studies and outlines how the studies lead to the forecast techniques used today.

The interest in Florida's weather is due to the frequency and regularity of its convective events. Even today, many of the thunderstorms are difficult to predict with a high degree of accuracy. The fact that Florida has the highest number of thunderstorms per year in the U.S. has been known for quite some time. In 1945, the United States Weather Bureau produced a report, using data as far back as 1906, showing Florida with

the highest average number of thunderstorm occurrences (Byers and Rodebush, 1948). Radar was used to produce an even finer scale picture of this frequency of occurrence (Frank *et al.* 1967). The variability of the frequency of thunderstorms, both spatially and temporally, has also been studied. Byers and Rodebush (1948), Gentry and Moore (1954), and Frank *et al.* (1967) looked at the organization of convective activity with respect to synoptic-scale forcing. These early studies tried to show that forecasting from purely a synoptic point of view was not sufficient.

Thermal mechanisms could not sufficiently explain how thunderstorms in Florida were created, in cases where no synoptic forcing was present, so a dynamic explanation was pursued. Byers and Rodebush (1948) pointed out that thermal instability is present continuously, but does not represent a mechanism sufficient to cause thunderstorms. They postulated that low-level horizontal convergence, due to the sea breeze, was the necessary ingredient for thunderstorm generation. Further studies began focusing on sea breezes.

By the early 1950s, the interaction of the sea breeze with the local environment was beginning to be explained with models and observations. Gentry and Moore (1954) used precipitation records to estimate the correlation of the sea breeze to the temporal and spatial location of convection. Estoque (1962) explained the interaction of the sea breeze with the synoptic wind field. He pointed out that if the synoptic flow is onshore before the onset of the sea breeze the resulting sea breeze is weaker because advection inhibits a large rise in air temperature. Frank *et al.* (1967) used radar data to show that the location of thunderstorms moves across the peninsula with the sea breeze. Sea breezes became the accepted reason for the formation of thunderstorms. Unfortunately, forecasting the sea breeze is as difficult as forecasting thunderstorms.

11

Continued study of the sea breeze refined the understanding of its structure and increased the knowledge of its mechanics. Reed (1979) differentiated between local scale sea breezes and peninsular scale sea breezes by measuring the strength of the diurnal oscillation of wind velocity. Nicholls *et al.* (1991) showed soil moisture to be an important factor in the speed of movement of the sea breeze front. More average soil moisture slows the speed of the sea breeze front, while less average moisture increases the speed. Xian and Pielke (1991) found the strength of the sea breeze circulation to be associated with the initial static stability and the rate of heating. Further study of the interaction of the sea breeze with the synoptic wind field by Bechtold *et al.* (1991) showed displacement of the sea breeze front in the direction of the synoptic flow.

Many of the recent thunderstorm studies focus on how the interaction between the sea breeze and the environment affects convection. Four scales of processes were suggested for the cause of thunderstorms in Florida (Cooper *et al.*,1982). These scales were synoptic, peninsular, mesoscale, and convective scale. Cooper described how the precursor conditions at each of the scales can determine the dynamics of the other scales. This creates a feedback loop which continues until surface heating stops at sunset. López *et al.* (1984) found that the vertical distribution of moisture is related to the degree of convective activity. More moisture in a layer up to 650mb correlated to more thunderstorms. Lopez *et al.* (1984) and Bauman *et al.* (1997) found that the stabilizing effect of warming and drying, caused by the subsidence from the Atlantic High reduced the chance of thunderstorms despite a strong sea breeze. These studies provided insight into when and where sea breezes form.

2.2   NPTI Prediction Technique

The work done by Charles Neumann in the late 1960's is the basis for the Neumann-Pfeffer Thunderstorm Index used today. Neumann published three reports, one of which became a journal article, which describe the specific thunderstorm environment at Cape Canaveral. In addition, the papers report on the factors found useful in forecasting thunderstorms. This section reviews those factors and describes the methods used to create the NPTI.

The first report by Neumann in 1968 provided a climatological study of thunderstorm occurrences in the Cape Canaveral area. It provided a detailed description of the frequency of thunderstorms. Neumann found the distinct shape of the probability distribution suggests that thunderstorm probability is dependent on what day of the summer it is. Days were numbered sequentially from May through September, and then the distribution was smoothed to more clearly show the trend. Neumann used a 15 day moving average to smooth the data (See Figure 3). The equation used to obtain this average, A, is given by equation 1,

$$A_n = \frac{1}{15} \cdot \sum_{k=n-7}^{n+7} T_k \qquad (1)$$

where n = day number,
    T_k = frequency of thunderstorm occurrence for that day number.

One discovery was the shape of the distribution of thunderstorm occurrence. This led to the day number being used as a regression variable to incorporate climatology into the forecast.

13
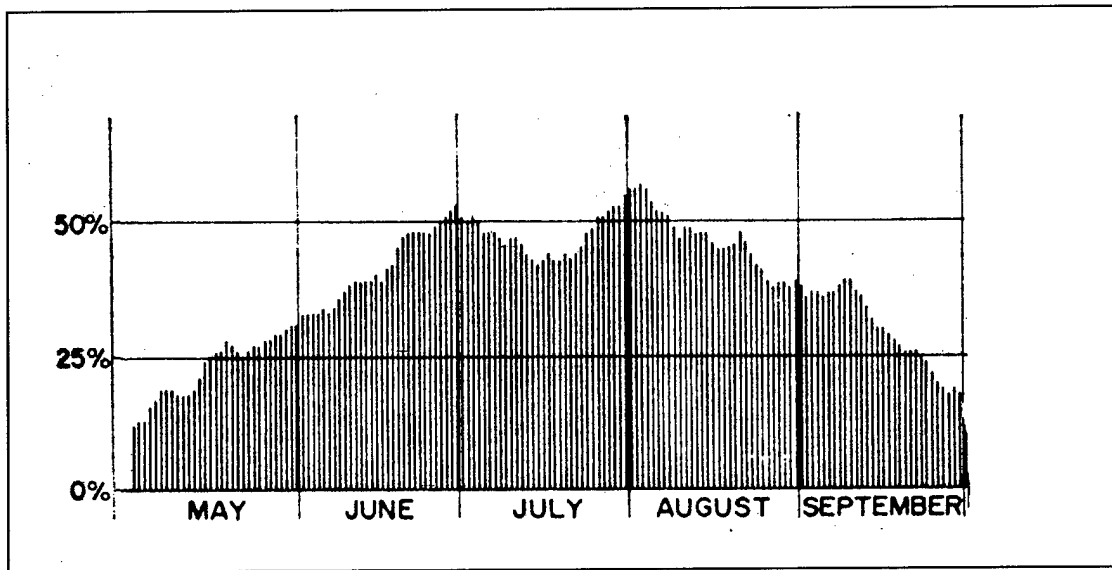
Figure 3 15-Day Average of Thunderstorm Frequency
(adapted from Neumann,1971)

The second report by Neumann in 1970 provided a climatological study of the

3000-foot winds in the Cape Canaveral area. Wind direction and speed were transformed

to u and v components. This avoided the discontinuity created as wind crossed from 360°

to 0° and allowed the wind to be converted to a scalar quantity. The 850-mb winds were

chosen to represent the 3000-foot winds used in the 1970 study. This level was used to

allow information derived from the sea breeze to be exploited as a predictor and was also

found to have a large correlation to thunderstorm occurrence. The 500-mb winds were

chosen to represent the upper level synoptic regime. This regime could indicate the

presence of subsidence. Neumann fitted the u and v components of the winds at both

850-mb and 500-mb, for each month, to a binomial distribution. These binomial

distributions were then used as the semi-major and semi-minor axes of an ellipse.

Neumann used the ellipses as functions to set limits on the values used in the regression.

This bounding was necessary because the Regression Estimation of Event Probabilities

(REEP) used to create a prediction equation can result in unrealistic probabilities for

14

extreme values of variables. The key result of his report was a better understanding of how local wind correlated with thunderstorms.

Finally, Neumann chose the remaining variables after calculating the correlation of many kinds of atmospheric parameters to thunderstorms. Two factors correlated the most to thunderstorm occurrence. They were the mean 800-mb to 600-mb relative humidity and the Showalter Stability Index (SSI). Both factors measure the stability of the atmosphere and were included as regression variables.

Neumann used two techniques to account for the nonlinear trends in the data. The first technique used second and third order polynomials to represent each independent variable in an initial regression. These polynomials allowed the coefficients to better fit a curve to each variable. The second technique was to regress the results of the first regression as opposed to regressing all the polynomials at once. The polynomial function used as an initial regression of the 850-mb winds is given by equation 2,

$$f(X_1) = f(s_{850}, t_{850}) = a_0 + a_1 s_{850} + a_2 t_{850} + a_3 s_{850} t_{850} + a_4 s_{850}^2 + a_5 t_{850}^2 \tag{2}$$
$$+ a_6 s_{850}^3 + a_7 s_{850}^2 t_{850} + a_8 s_{850} t_{850}^2 + a_9 t_{850}^3$$

where X_1 = 850-mb winds,
        s = the u wind component in kt,
        t = the v wind component in kt,
        $a_{0..9}$ = the regression coefficients.

The polynomial function for the initial regression of the 500-mb winds is given by equation 3,

$$f(X_2) = f(u_{500}, v_{500}) = b_0 + b_1 u_{500} + b_2 v_{500} + b_3 u_{500} v_{500} + b_4 u_{500}^2 + b_5 v_{500}^2 \tag{3}$$
$$+ b_6 u_{500}^3 + b_7 u_{500}^2 v_{500} + b_8 u_{500} v_{500}^2 + b_9 v_{500}^3$$

where X_2 = 500-mb winds,
        u and v = the u and v wind components in kt,
        $b_{0..9}$ = the regression coefficients.

The polynomial functions for 800-mb to 600-mb mean Relative Humidity, SSI, and day number are given by equations 4, 5, and 6 respectively,

$$f(X_3) = c_0 + c_1X_3 + c_2X_3^2 + c_3X_3^3 \qquad (4)$$

$$f(X_4) = d_0 + d_1X_4 + d_2X_4^2 \qquad (5)$$

$$f(X_5) = e_0 + e_1X_5 + e_2X_5^2 \qquad (6)$$

where X_3 = Mean relative humidity, 800-mb to 600-mb in percent,
X_4 = Showalter Stability Index in degrees Celsius,
X_5 = Day number,
$c_{0..3}$, $d_{0..2}$, and $e_{0..2}$ = regression coefficients.

The combination of the results of the previous regressions used in a final regression is given by equation 7,

$$P = g_0 + g_1f(X_1) + g_2f(X_2) + g_3f(X_3) + g_4f(X_4) + g_5f(X_5) \qquad (7)$$

where P = Probability of thunderstorm occurrence,
$g_{0..5}$ = the regression coefficients,
$X_1..X_5$ = the results of the previous functions

The final result represents the probability of thunderstorms for that afternoon. Using the REEP method, any results greater than 1 or less than 0 are rounded off to fit within that interval.

2.3   Conditional Climatological Frequency

The probability for the occurrence of a thunderstorm can also be calculated from the historical sequence of thunderstorm events. To understand this method of forecasting, it is important to understand how a historical sequence is related to conditional climatology and persistence. Once this relationship is understood, a forecast

16

using the sequence can be derived. This forecast can be thought of as a kind of multiple-day persistence.

Conditional climatology can be used as an alternative to probability derived from simple relative frequency. The method for computing both is similar. As an example, the relative frequency of thunderstorm days can be calculated from the ratio of days thunderstorms occurred to the total number of days observed. Similarly, the conditional probability is the ratio of the number of days with thunderstorms and some other event to the total number of days with thunderstorms. In this case, the conditional probability is a subset of the relative frequency. Using the conditional probability gives a more accurate forecast; however, the parameter used as a condition must be statistically related to the occurrence, or the conditional probabilities will be the same with or without the parameter (Wilks, 1995). This leads to the consideration of persistence as a possible parameter.

Persistence should be understood before being used as a forecast variable. Statistically, persistence describes the dependence among successive events. For that reason, another name for persistence is positive serial dependence (Wilks, 1995). Using this concept, if a thunderstorm occurred yesterday, the probability for a thunderstorm today is higher. This dependence on prior events implies that the second event is conditional to the first with a partial cause-and-effect relationship. Therefore, persistence is a conditional probability (Wilks, 1995).

For weather forecasting, the conditional probability for persistence is often rounded to 0% or 100%; however, an exact value can be calculated using climatological relative frequencies. This calculation is made from a set of recorded prior events. For

17

persistence, two climatological relative frequencies are determined: the relative frequency of thunderstorms given no thunderstorms yesterday, and the relative frequency of thunderstorms given thunderstorms did occur yesterday. The fact that the frequencies are found not to be equal shows there is dependence between the occurrence of thunderstorms yesterday and thunderstorms today. The magnitude of the difference between the two probabilities is a measure of the strength of dependence (Wilks,1995). It is the strength of the dependence that is exploited to improve upon the persistence method of forecasting.

For this study, determining the conditional probability given multiple days of occurrence or nonoccurrence required recording the pattern of events. The method created uses a sequential record, spanning a specific number of days, showing the pattern of when a thunderstorm did or did not occur. A day with a thunderstorm occurrence is labeled "1," and a day without a thunderstorm occurrence is labeled "0." Ordered with the most recent day on the left, thunderstorms yesterday but no thunderstorms the day before would be recorded as "10." This represents a 2-day pattern of persistence. Six days might appear as "110010." It is these patterns that become the condition for thunderstorms. For persistence, if the equation for the conditional probability (p) of thunderstorm occurrence today given a thunderstorm occurrence yesterday is

$$p1_1 = \Pr\{X_t = 1 \mid X_{t-1} = 1\} \quad , \qquad (8)$$

then conditional probability for two day persistence is

$$p0_{00} = \Pr\{X_t = 0 \mid X_{t-1} = 0, X_{t-2} = 0\} \qquad (9a)$$

$$p0_{01} = \Pr\{X_t = 0 \mid X_{t-1} = 0, X_{t-2} = 1\} \qquad (9b)$$

$$p1_{10} = \Pr\{X_t = 1 \mid X_{t-1} = 1, X_{t-2} = 0\} \qquad (9c)$$

$$p1_{11} = \Pr\{X_t = 1 \mid X_{t-1} = 1, X_{t-2} = 1\}. \qquad (9d)$$

These conditional probabilities are also known as transition probabilities in a first order

Markov Chain (Wilks, 1995). The multi-order conditional probabilities are found by

expanding the first-order Markov chain,

$$p1_{ij} = \Pr\{X_t = 1 \mid X_{t-1} = i, X_{t-2} = j\}$$
$$\vdots \qquad (10)$$
$$p1_{ijk...x} = \Pr\{X_t = 1 \mid X_{t-1} = i, X_{t-2} = j, X_{t-3} = k, \ldots\ldots, X_{t-m} = x\}$$

The variables i,j,k,..x represent the historical sequence of thunderstorm occurrence. The

probability calculated in this manner would represent the probability of a thunderstorm

given a pattern of occurrence in the past. This probability could be used as the forecast as

opposed to persistence's 0% or 100% forecast.

The conditional nature of thunderstorms is expected given the assumption that the

environment causing the thunderstorm does not change or move from day to day, and

synoptic changes are rare or occur after long time periods. This leads to the question of

how far back the dependence is statistically significant. The occurrence of thunderstorms

may be dependent on not only the previous day but also on days prior. The next step is to

find the conditional probability given different patterns of occurrence over a longer

period and test the statistical significance. Presumably, the farthest time period back that

can be used is the time scale of synoptic changes. At this time, thunderstorm occurrence

is no longer dependent on prior occurrences, but on some synoptic change.

Unfortunately, the availability of data often restricts the scope of the test.

19

2.4 Regression Methods

There are many methods for finding functions to fit curves to data, and the worth of each is dependent on the application. Constraints such as availability of data, type of data, and type of response dictate which regression method to use. This section gives a further description of the regression method used by Neumann and explains the use of logistic regression as an alternative.

2.4.1 Multiple Multivariate Linear Regression

The Regression Estimation of Event Probabilities (REEP) is a regression method often used in meteorology to find a function with prediction capabilities. The results of the function created by this method of regression are considered to be an approximation of probability of occurrence (Wilks, 1995). Unfortunately, unrealistic probabilities can

Level 1 Regression

$$y_1 = \boxed{A_0} + \boxed{A_1} x_1 + \boxed{A_2} (x_1)^2$$

$$y_2 = \boxed{B_0} + \boxed{B_1} x_2 + \boxed{B_2} (x_2)^2$$

Level 2 Regression

$$P = \boxed{C_0} + \boxed{C_1} y_1 + \boxed{C_2} y_2$$
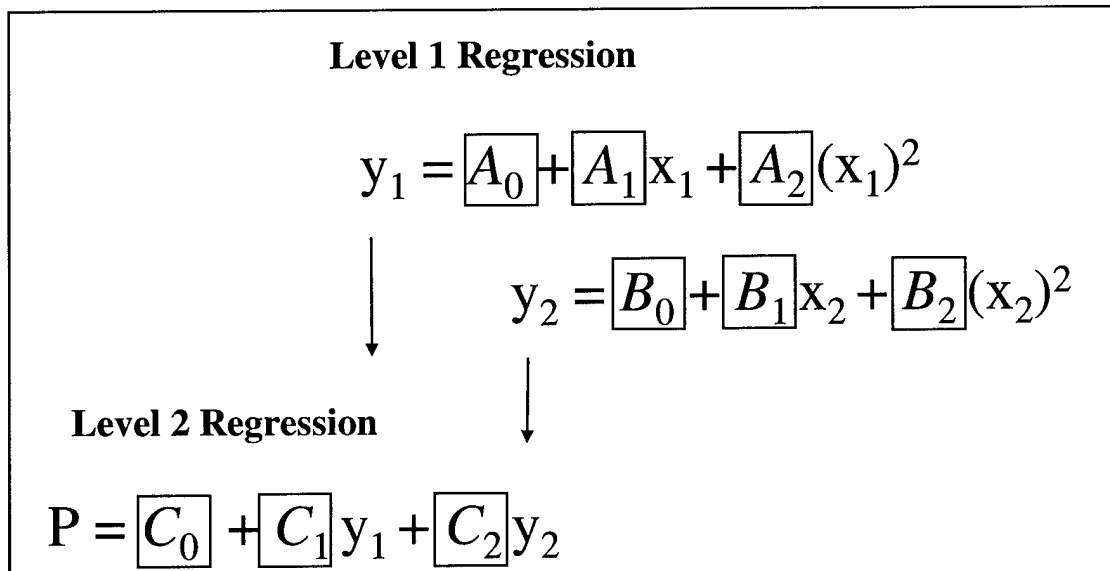
Figure 4 Multiple Multivariate Regression Functions

result when values outside those originally regressed are used for prediction. In the event a result is found to be above 1 or below 0, the probability is assumed to be 1 or 0, respectively. This regression method was used by Neumann to find a set of prediction algorithms.

Multiple multivariate linear regression is an extension of REEP that takes into account some forms of nonlinear response. To do this, the regression is divided into two levels. In the first level, a polynomial function of each variable is created. This is done to include the nonlinear response of each parameter in the regression. Then each function is linearly regressed against thunderstorm occurrence. The outcome for each regression in level one is a set of coefficients for each polynomial function. These coefficients are used with their corresponding function to calculate a new value. The values calculated from each function are passed to level two where they are the variables for a new multivariate function (Figure 4). When the level two function is regressed against thunderstorm occurrence, another set of coefficients is created. These level two coefficients are the slope and intercepts of the surfaces represented by the functions. These coefficients are used with the level two function to calculate the probability of a thunderstorm. Once all the regressions are complete and the coefficients known, the procedure to find the probability of a thunderstorm is relatively simple. First, enter the meteorological variable from independent data into its corresponding level one function. Insert the results from all the level one functions into the corresponding variables in the level two function. The result of this function represents the probability of thunderstorm occurrence and is continuous from 0 to 1.

Continuous Linear Response

Occur (1)

Dichotomous Observations

Not Occur (0)

$$y_1 = B_0 + B_1 \, x_1 + \ldots$$

Continuous Logistic Response

(1)

Dichotomous Observations

(0)

$$y_2 = \dfrac{e^{(B_0 + B_1 x_1 + \ldots)}}{1 + e^{(B_0 + B_1 x_1 + \ldots)}}$$
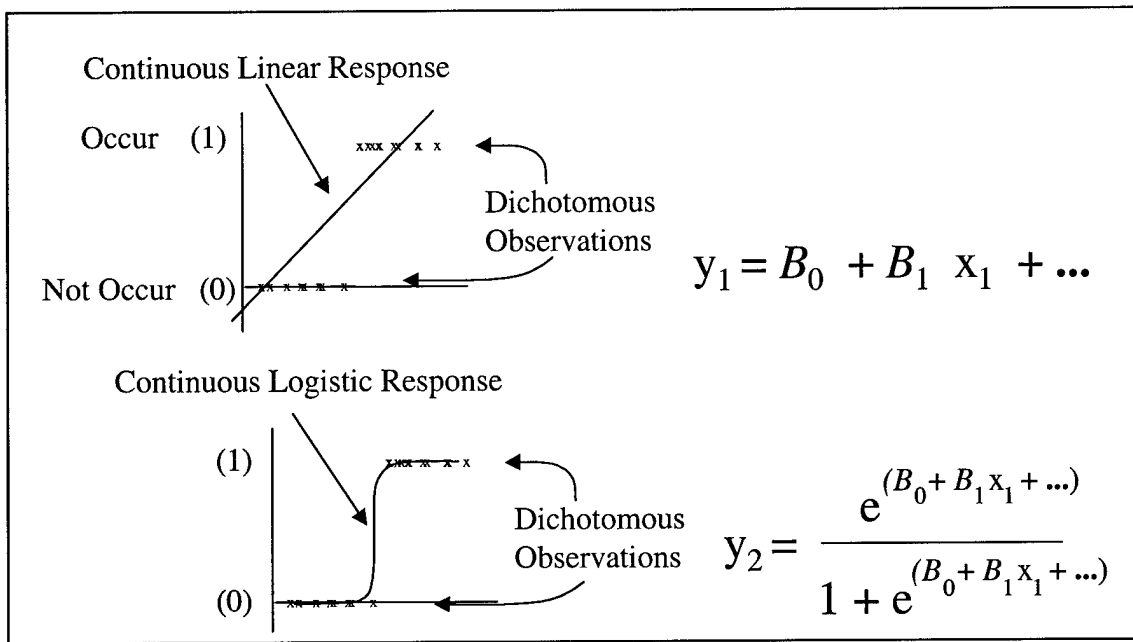
Figure 5 Examples of Regression Responses

2.4.2   Logistic Regression

Logistic regression is a more exact method of modeling nonlinear dichotomous weather events. When the dependent variable is not continuous, the response function should be curvilinear (Neter *et al.*, 1983). To meet this constraint, logistic regression maps the data to an asymptotic logarithmic distribution. Accordingly, logistic regression has the advantage of being automatically constrained to a range of probabilities between 0 and 1 (Figure 5). This characteristic is due to the asymptotic nature of the exponential function used as the model distribution. The distribution modeled in Figure 5 is given by equation 11,

$$E(Y) = \frac{e^{\left(\beta_0 + \beta_1 X\right)}}{1 + e^{\left(\beta_0 + \beta_1 X\right)}} \tag{11}$$

where β_0 = a coefficient that affects placement of the curve laterally,

β_1 = a coefficient that alters the slope of the curve (Dillon and Goldstein, 1984).

Multiple multivariate logistic regression follows the same steps as multiple multivariate linear regression to arrive at a predictive function.

2.5   Statistical Forecast Verification

Statistical analysis is a necessary part of any forecasting process because no forecasting method is perfect. To estimate the effectiveness of a forecasting tool, some measure of accuracy must be used. This section describes the statistical measures used in this research, which are common in the field of weather forecasting.

The 2 X 2 contingency table is a simple method of displaying all outcomes for a binomial process. Figure 6 shows how each quadrant is defined and gives the calculations for marginal frequencies. N represents sample size. For a perfect forecast, b

Figure 6 Two-By-Two Contingency Table

23

and c would contain entries of 0 with a + d = N (Wilks, 1995). For example, each time a thunderstorm was forecast it occurred, and each time no thunderstorm was forecast none occurred. The numbers entered in each block of a contingency table are dependent on a yes or no forecast. However, most forecasts are a probability, not yes or no. To use the contingency table it is necessary to define a probability above which the event is forecast to occur and below which the event is forecast not to occur. This probability is called a cutoff value. Once the cutoff value is decided, the probability forecasts can be categorized as yes or no forecasts, and the contingency table can be completed. It is important to understand that the accuracy of the forecast and the values for each block of the contingency table are highly dependent on the choice of cutoff value.

The contingency table can be used to calculate various measures of accuracy. Hit rate (HR), given by equation 12, calculates the percent of correct forecasts for the total number of forecasts and is the most easily understood and widely used (Wilks, 1995):

$$HR = \frac{a+d}{N} \cdot 100 \; .$$

(12)

It is calculated from the contingency table and describes the general quality of the method. For thunderstorm forecasting, it describes how often a forecast is correct. The best hit rate is 100% and the worst is -100%.

Threat score (TS), given by equation 13,

$$TS = \frac{a}{a+b+c} \cdot 100$$

(13)

measures the percent of correct forecasts for the number of times an event occurred or

was forecast to occur (Wilks, 1995). Threat Score, sometimes called critical success index (CSI), does not have some of the undesirable qualities of hit rate. Hit rate describes the ability of a method to get all forecasts correct while TS calculates the ability of a method to get all forecasts correct after removing correct "no thunderstorm" forecasts. TS does not give credit for forecasting a thunderstorm not to occur but it counts against the method if it,s wrong.

Sometimes it is beneficial to know the accuracy of a "no thunderstorm" forecast as well. This can be provided by using Threat Score No (TSN). TSN calculates the ability of a method to get all forecasts correct after removing correct "yes thunderstorm" forecasts. Equation 14 calculates the percent of correct "no thunderstorm" forecasts after removing correct "yes thunderstorm" forecasts:

$$TSN = \frac{d}{b+c+d} \cdot 100 \quad .$$

(14)

The best threat score is 100% and the worst 0%.

The Probability of Detection (POD), given by equation 15,

$$POD = \frac{a}{a+c} \cdot 100$$

(15)

measures the ratio of correct "yes thunderstorm" forecasts to the number of times a thunderstorm occurred (Wilks, 1995). The best POD is 100% and the worst is 0%.

The Probability of Detection No (PODN), given by equation 16,

$$PODN = \frac{d}{b+d} \cdot 100$$

(16)

measures the ratio of correct "no thunderstorm" forecasts to the number of times a thunderstorm did not occur (Wilks, 1995). The best PODN is 100% and the worst is 0%.

The False Alarm Rate (FAR), given by equation 17,

$$FAR = \frac{b}{a+b} \cdot 100 \tag{17}$$

measures how often an incorrect "yes thunderstorm" forecast was made versus the number of times "yes thunderstorm" was forecast (Wilks, 1995). The best FAR is 0% and the worst is 100%. The False Alarm Rate No (FARN), given by equation 18,

$$FARN = \frac{c}{c+d} \cdot 100 \tag{18}$$

measures how often an incorrect "no thunderstorm" forecast was made versus the number of times "no thunderstorm" was forecast (Wilks, 1995). The best FARN is 0% and the worst is 100%.

The Skill Score (SS), is given by equation 19,

$$SS_{ref} = \frac{A - A_{ref}}{A_{perf} - A_{ref}} \cdot 100 \tag{19}$$

where A = measure of accuracy for the method of interest,
A_{ref} = measure of accuracy for the method used as reference (usually persistence),
A_{perf} = measure of accuracy for a perfect forecast method.

SS compares a forecast method to a reference forecast method such as persistence. It measures how much better or worse than the reference forecast the forecast of interest is. The best SS is 100% and the worst is 0%.

The Bias Ratio, given by equation 20,

$$BIAS = \frac{a + b}{a + c}$$ (20)

shows whether a forecast method is over forecasting or under forecasting an event. Over forecasting creates a bias greater than 1 while under forecasting creates a bias less than 1. A perfect forecast has a bias of 1.

The Heidke Skill Score (HSS) compares the hit rate of the forecast method to the hit rate from a random forecast (Wilks, 1995). Given by equation 21,

$$HSS = \frac{2(ad - bc)}{(a + c)(c + d) + (a + b)(b + d)} .$$ (21)

HSS is a simplified skill score which uses as a reference forecast method the hit rate of a random forecast with the same marginal distributions as the verification data set. The hit rate from a random forecast is calculated from the marginal probabilities in the contingency table of the verification data set. A perfect score is 1.

The Kuipers Skill Score (KSS) is similar to HSS. It compares the hit rate of two forecast methods. Given by equation 22,

$$KSS = \frac{ad - bc}{(a + c)(b + d)} .$$ (22)

KSS is also a skill score, but it uses the hit rate of an unbiased random forecast for the reference forecast method (Wilks, 1995). A perfect score is 1 and the worst score is -1.

The Brier Score (BS), is given by equation 23,

$$BS = \frac{1}{n} \cdot \sum_{k=1}^{n} (y_k - o_k)^2$$

(23)

where n = number of cases,
       y_k = forecast probability,
       o_k = observed probability (1 if occurred, 0 if did not occur).

The BS measures the accuracy of a probabilistic forecast for a dichotomous event (Wilks, 1995). It has the advantage of calculating a measure of accuracy without first specifying a cutoff value. The Brier Score is equivalent to the mean-squared error. A perfect Brier Score is 0 and the worst is 1.

The Ratio Skill Score (RSS) shows how one forecast method compares to a reference forecast method. It compares the Brier Score of the forecast method of interest to the Brier Score of the reference forecast method. RSS, given by equation 24,

$$RSS = \frac{BS_{ref} - BS}{BS_{ref}} \cdot 100$$

(24)

where BS = the Brier Score of the forecast method of interest,
      BS_{ref} = the Brier Score of the reference forecast.

RSS is the same as the skill score with a perfect forecast of 0.

To show that the result of the forecast is dependent on the forecast and is not random, a χ^2 test is performed. This gives the forecast validity and shows that the result

is statistically significant. First, the expected value for each block of the contingency table is calculated using equation 25,

$$E_{ij} = \frac{n_i \cdot n_j}{N}$$

(25)

where E_{ij} = the expected value of the i,j block of the contingency table,
N = total sum,
n_i = marginal sum of the ith row,
n_j = marginal sum of the jth column.

The expected value is used in the Pearson's Chi Squared Test to find χ^2 in equation 26,

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{\left(n_{ij} - E_{ij}\right)^2}{E_{ij}}$$

(26)

where E_{ij} = the expected value of the i,j block of the contingency table,
n_{ij} = the value of the i,j block of the contingency table.

The χ^2 is compared to its distribution using one degree of freedom (Everitt, 1992). χ^2 greater 3.843 shows there is dependence between the forecasts and observations at the 95% confidence level. The larger χ^2 is the stronger the dependence (Devore, 1995).

29

3. *Methodology*

This chapter describes the several lengthy steps involved in improving the NPTI. The first section describes methods of data organization and the data manipulation used to obtain several arrays of observations. The second section describes how additional variables were derived from the basic information provided by the original data. The third section describes the analysis of the data to determine the best variables to be regressed. The fourth section explains how the data was regressed to create algorithms, and the last section explains how the various methods were verified.

3.1 *Data*

It is important to understand the type and makeup of the data used for this research, and to be aware of how it was manipulated. This allows for an understanding of what the final algorithm represents and how it can be affected by various changes in the variables. The data were obtained through several different sources, combined, and reorganized to allow division into different categories. Then the data were placed into an array and, where possible, missing values were interpolated. The first subsection describes where the data comes from and its content and format. The second subsection describes how the data were modified and what the benefit was. The last subsection describes how the data were organized.

Table 1 Surface Observations Available

	May	June	July	August	September
Years with Observations	1950 – 1977 1982 1987 - 1996	1950 – 1977 1982 1987 - 1996	1950 – 1977 1982 1987 - 1996	1950 – 1977 1982 1987 - 1996	1950 – 1977 1982 1987 - 1996

3.1.1   Data Acquisition

Surface observations were obtained from the Air Force Combat Climatology Center (AFCCC). The observations contained the standard meteorological parameters such as temperature, pressure, winds, humidity, weather description, and visibility. Table 1 shows which years had observations available for each month. The surface observations were stored as a single text file with each line of text containing all the elements of a single observation. Because the valid time of the forecast is from 1100 UTC to 0400 UTC the next day (0700 hours to 2400 hours Eastern Standard Time), only thunderstorms documented within this time period were considered. Each observation was screened for the acronyms TS, TSRA, +TSRA, and –TSRA for each day from May to September of each year. A day with a thunderstorm occurrence was annotated with a 1, while a day with no thunderstorm occurrence was annotated with a 0. Appendix A contains the Fortran code for screening the surface observations.

Upper air observations taken by rawinsonde were also obtained from the AFCCC during a previous research effort. Not every day had an observation and some years had no observations at all. A large gap in data occurred from 1971 to 1982.

31

Table 2 Upper Air Observations Available

	May	June	July	August	September
Years with Observations (including missing data)	1950 – 1970 1983 – 1985 1987, 1988 1991 - 1995	1950 – 1969 1983 – 1985 1987, 1988 1991 - 1995	1950 – 1969 1983 – 1989 1992 - 1995	1950 – 1969 1983 1985 - 1988 1992 - 1995	1950 – 1969 1983 – 1988 1992 - 1995
Years with Observations (after missing data removed)	1950 – 1952 1954 1956 - 1970 1983 – 1985 1987, 1988 1994	1950 – 1953 1955 - 1969 1983 – 1985 1987, 1988	1950 – 1969 1983 – 1988	1950 – 1969 1983 1985 – 1988	1950 – 1969 1983 – 1988

Table 2 shows which years had upper air observations available for each month. The upper air observations were stored as a single text file, and each line of text represented one pressure level of a specific sounding. The elements recorded for each pressure level include height, temperature, dew point, wind direction, wind speed, and relative humidity. Some elements were missing for various pressure levels, and every sounding measured different pressure levels in addition to the mandatory pressure levels. In addition, the hours soundings were available were different from day to day. The NPTI algorithm forecasts thunderstorms for the day using information available from that morning. Therefore, only data within the interval from 0900 UTC to 1600 UTC were used. The set of observations from each hour was saved as a separate file. This allowed later divisions of data to be made more easily. Appendix B contains the Fortran code for screening the upper air observations.

Because multiple years of surface observation records were missing, records of lightning strikes were used as a substitute to document thunderstorm occurrence. See

Table 1 for missing years. Lightning strike records from 1985 to 1997 were obtained, courtesy of Dr. Richard E. Orville, from Global Atmospherics Corporation. Of the years of lightning strike data available, only 1985 was missing surface observations. The data are a record of flash days, with a count of how many cloud-to-ground lightning strikes occur per day within a designated set of coordinates. Any day with a recorded lightning strike was interpreted as a thunderstorm day and annotated with a 1. Otherwise, it was assumed no thunderstorm occurred and the day was annotated with a 0. To validate this assumption, days with both with surface observations and lightning strike data were compared. The area bounded by the coordinates 28.2° N to 28.6° N and -80.3° E to -80.75° E resulted in the least error, less than 10%, between the two data sets and was used for the year without surface observations, 1985. Thunderstorm days transformed from lightning strike data were incorporated into the climatological frequency of thunderstorms derived from surface observations. With this data set, an extra year, 153 days, of information were added to the data available to be analyzed.

3.1.2   Data Manipulation

It is not uncommon for data to be missing from an upper air sounding. Unfortunately, to forecast using NPTI, every variable must have a value, or a result for the algorithm cannot be calculated. This means only soundings with all the required values present can be used. Therefore, it is important to ensure that as much missing data can be accounted for as possible.

If the value for an element of a sounding is missing, the whole sounding does not necessarily need to be discounted. Even if the missing data are from a necessary level,

the surrounding data may provide a clue as to what the value for the missing piece of data might be. Assuming temperature and dew point are continuous across small distances, it is possible to account for missing data by interpolation. The accuracy of interpolation depends on how much data are missing and how close the nearest known data point is. The closer the pressure level of a known element is to the pressure level of the missing element, the more likely the interpolation will be accurate.

To check this method of data replacement, data denial verification was accomplished. Soundings with no measurements missing were selected and saved in a separate data set. Those soundings then had measurements purposely deleted (withheld). Values were estimated for the deleted values using interpolation, and the difference was measured between the true (withheld) value and the interpolated value. Five hundred thirty-four soundings with no data missing below 300-mb were used in the verification. Also, 9 interpolation methods were tested for filling in missing data. These included:

Linear interpolation	Linear with climatology
Cubic spline interpolation	Cubic spline with climatology
Averaging method	Averaging method with climatology
Minimum change method	Minimum change method with climatology
Temporal averaging	

Having a minimum mean square error was one condition for choosing which method to use. While the mean square error was lowest for cubic spline methods, it produced some extreme outlying errors. Figure 7 shows an example of when a method with a higher mean square error is better than a method with extreme outlying errors. These outlying errors represent points at which the interpolation resulted in large over- or under-estimates for a value. Thus, minimizing the error at outlying data points (outliers) was

34

Table 3 Mean Square Errors of Interpolation

Location of Missing Data	Lowest Mean Square Error		Highest Mean Square Error	
1 point missing center of data set	Temperature Dew point	0.452 °C 1.473 °C	Temperature Dew point	0.873 °C 2.452 °C
1 point missing end of data set	Temperature Dew point	0.693 °C 0.888 °C	Temperature Dew point	3.388 °C 1.162 °C
7 points missing	Temperature Dew point	0.660 °C 1.610 °C	Temperature Dew point	3.398 °C 4.676 °C
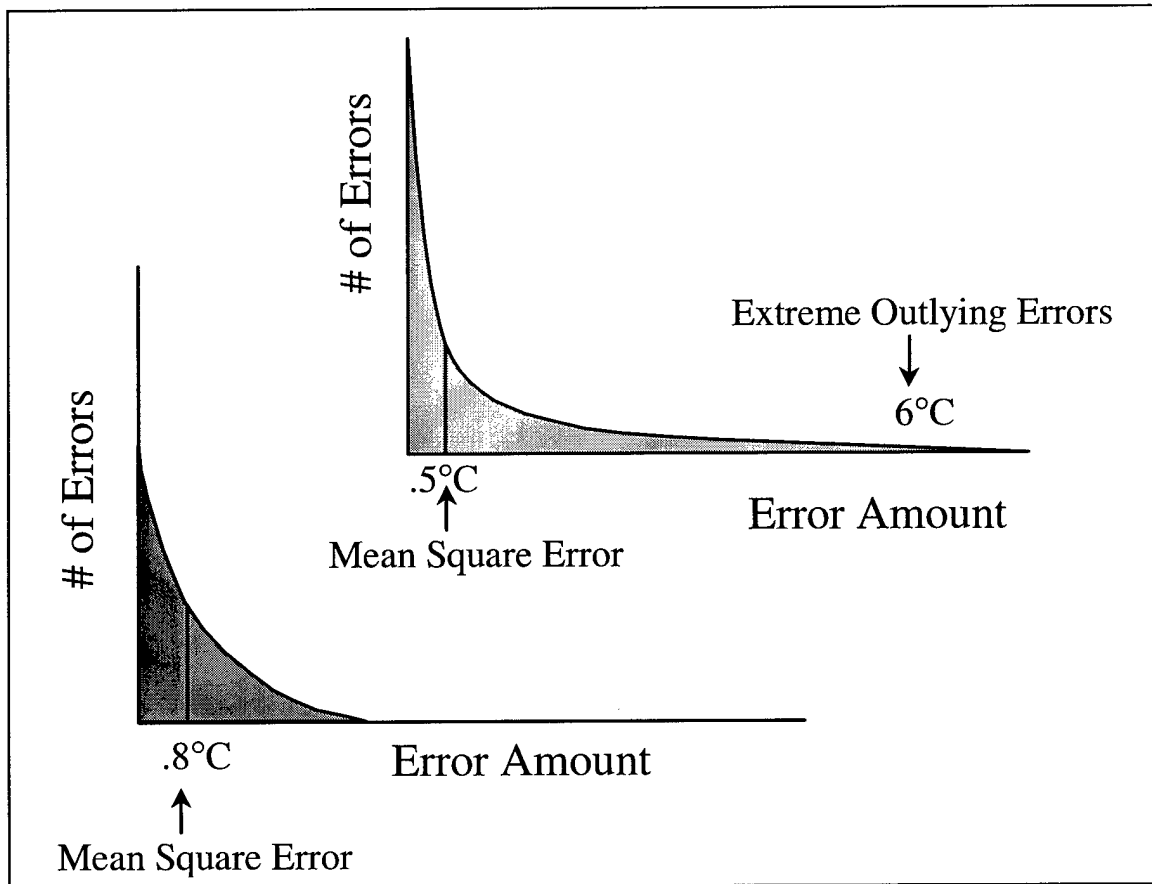
Figure 7 Outlying Errors

the second condition for choosing an interpolation method. Table 3 shows where the highest and lowest mean square errors occurred.

It was found that interpolation accuracy also depended on the amount of data missing and on where within the observation the missing data occurred. In general, cubic spline interpolation was the most accurate if few points were missing. The more measured data points available, the more accurate the determination of the 1st and 2nd derivatives were. An accurate determination of the 1st and 2nd derivatives allows a more accurate estimate of a missing point. However, cubic spline interpolation also had some of the largest outliers. Linear interpolation was found to be more accurate than other methods of interpolation if many data points were missing or data were missing at the ends of the data set. Linear interpolation also had the fewest outliers. The characteristics of these two methods led to the creation of the minimum difference and averaging methods. The minimum difference method calculates both the cubic spline and the linear interpolations and uses the value which has changed the least from the closest known value. The averaging method uses the average of the cubic spline and linear interpolations. Using the minimum difference method resulted in a lower mean square error than linear interpolation and a lower number of outliers than cubic spline interpolation alone. Thus, the minimum difference method was chosen to recover missing data during the rest of the research.

Two factors were considered for implementation of an interpolation method. The first factor was deciding how much data to interpolate. With interpolation, any number of points can be chosen; however, too many would be impractical. Due to memory

Table 4 Observations Available for Final Data Set

	May	June	July	August	September
Final Interpolated Data set	1950 – 1970 1985 1987, 1988 1994	1950 – 1969 1985 1987, 1988	1950 – 1969 1985 – 1988 1994	1950 – 1969 1985 - 1988 1994	1950 – 1969 1985– 1988 1994

and disk space limitations, elements of pressure levels from 1000mb to 500mb in 50-mb

increments were chosen as the data points to be used. These levels also represent the

mandatory data levels, except 925-mb, from the standard upper air rawinsonde. A second

consideration in choosing which interpolation method to use was deciding how many

missing points would be allowed before the observation was considered unusable. For

linear interpolation, at least two other known points are needed, and three points are

needed for cubic spline interpolation. To ensure a minimum of error, only observations

with at least one data point below 800-mb, one data point between 600-mb and 800-mb,

and one data point above 600-mb were used for interpolation.

After all interpolation was complete and the days where interpolation could not be

used were removed, 61 temperature interpolations and 1528 dewpoint interpolations were

used to fill in missing data. This resulted in the recovery of 966 days, approximately 6

years, of upper air observations. Table 4 shows the observations available after

interpolation.

3.1.3   Data Organization

To keep track of observations and when they occur, a simple numbering system

was created. This was done because observations were not continuous throughout the

year, and the number of days per month is different depending on the month. Because

the first day of interest is May 1st and September 30th is the last day of interest, days of

summer were labeled 1 through 153. Each observation was assigned a day number

corresponding to when during the summer the observation was taken. Observations were

also assigned a number from 1 to 7191 which represented continuous numbering from

1950 to 1996 excluding October to April. May 1st 1951 was numbered 154. This

numbering system allowed upper air observations and surface observations to be easily

combined.

Once the upper air observations had been interpolated they were reorganized so

that each observation filled a single row in an array. Additional information, such as the

number of interpolations per observation and the number of known data points per

observation, were also recorded for use later in the study. The rows could then be

matched with the corresponding information about thunderstorm occurrence. With 6

elements per pressure level and 11 pressure levels, 66 columns of upper air

meteorological data were recorded. The end product was one array of numbers with rows

representing an observation and columns representing an element type such as

temperature or wind direction.

3.2   Derived Variable Calculation

With the measured data obtained from the observations, an additional three

categories of data can be derived. These categories are thunderstorm indices,

climatological frequency of thunderstorms, and multi-day persistence. Values for all

three categories were calculated and incorporated into the data set to be regressed. The

first sub-section below lists the thunderstorm indices used, and the following two sub-sections explain how the climatological frequency and multi-day persistence patterns were derived.

3.2.1 Thunderstorm Indices

Thunderstorm indices are created from a combination of meteorological variables and are meant to show the likelihood of a thunderstorm under the given meteorological conditions. A number of other parameters are calculated in an effort to insure all pieces of information with some possible predictive value were used. Equations for the these calculations were taken from the Air Weather Service (AWS) equation and algorithm guide (Duffield and Nastrom, 1983). Table 5 lists the thunderstorm indices and parameters computed for every observation and tested for possible use in the final algorithm.

Table 5 Variables Derived From Upper Air Observations

Thunderstorm Indices	Other Meteorological Parameters
Showalter Stability Index (SSI)	Lifted Condensation Level (LCL)
Lifted Index (LI)	Convective Condensation Level (CCL)
K Index	Mean Relative Humidity (1000-mb to 700-mb)
Vertical Totals (VT)	Mean Relative Humidity (600-mb to 800-mb)
Cross Totals (CT)	500-mb Thickness
Total Totals (TT)	700-mb Thickness
Thompson Index (TI)	850-mb Thickness
Microburst Day Probability Index (MDPI)	850-mb to 500-mb Thickness
	850-mb to 700-mb Thickness
	700-mb to 500-mb Thickness
	1000-mb to 500-mb Vertical Wind Shear
	700-mb to 500-mb Vertical Wind Shear
	1000-mb to 700-mb Vertical Wind Shear
	850-mb to 600-mb Vertical Wind Shear

40

3.2.2 *Frequency Analysis*

Climatological frequency represents how often, over a specified number of years, a day has had thunderstorms. The frequency can also be calculated for other time periods such as weeks, months, or years. An example of a daily climatological frequency can be seen in Figure 8 which comes from the surface observations described in section 3.1.1. Each bar represents the total number of thunderstorm days that occurred on that day of the summer over a 38 year period. By dividing the total number of thunderstorms by 38, the climatic probability can be calculated. Converting and using this frequency as a probability assumes that the probability of a thunderstorm is in some part dependent on the time of year. Because there is so much variability from day to day in the frequency of thunderstorms, the question becomes: how small a time frame can a climatological probability represent and still be useful? Neumann used a 15-day moving average but claimed that a shorter moving average might be better (Neumann, 1968). By calculating the climatological probability for a number of time periods and using an independent data set to verify each one, the optimum time period can be determined.
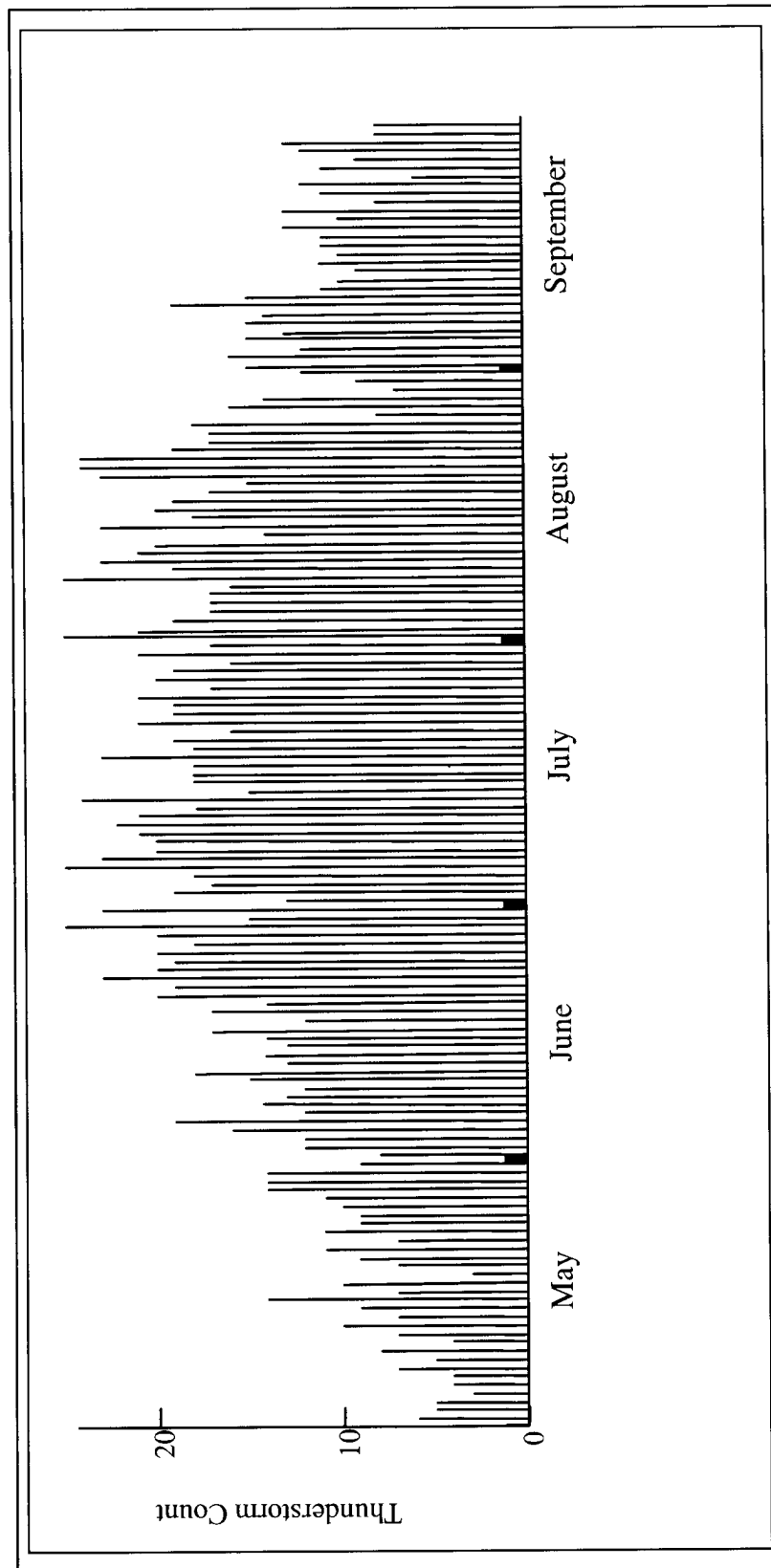
Figure 8 Thunderstorm Frequency Distribution

43
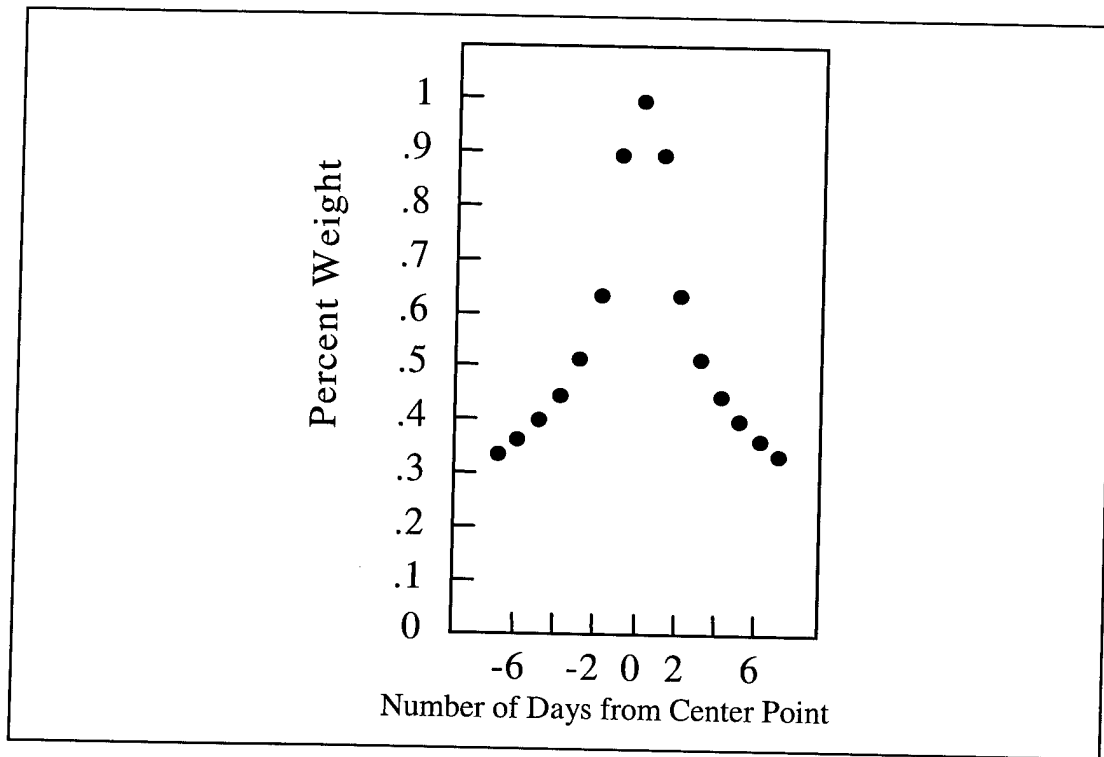
Figure 9 Weight Function for Climatological Frequency

For this study, a 15-day weighted moving average was chosen over a linear

moving average. The probability using a 15-day weighted moving average was obtained

using the equation 27,

$$P_n = \frac{1}{N} \cdot \frac{\sum_{k=1}^{7}\left[\frac{.9}{\sqrt{k}} \cdot \left(T_{n-k} + T_{n+k}\right)\right] + T_n}{\sum_{k=1}^{7}\left[\frac{.9}{\sqrt{k}} \cdot 2\right] + 1}$$ (27)

where P_n = weighted moving average on the day number of interest
    n = day number of interest,
    T = frequency on day n,
    k = number of days distant from day n,
    N = total number of years averaged.

Each day, except the day of interest, is weighted by $.9/\sqrt{k}$. The weighted days are

summed, and the total is divided by the sum of the weights and by N. This method gives

a decreasing weight to each day as shown in Figure 9.

44
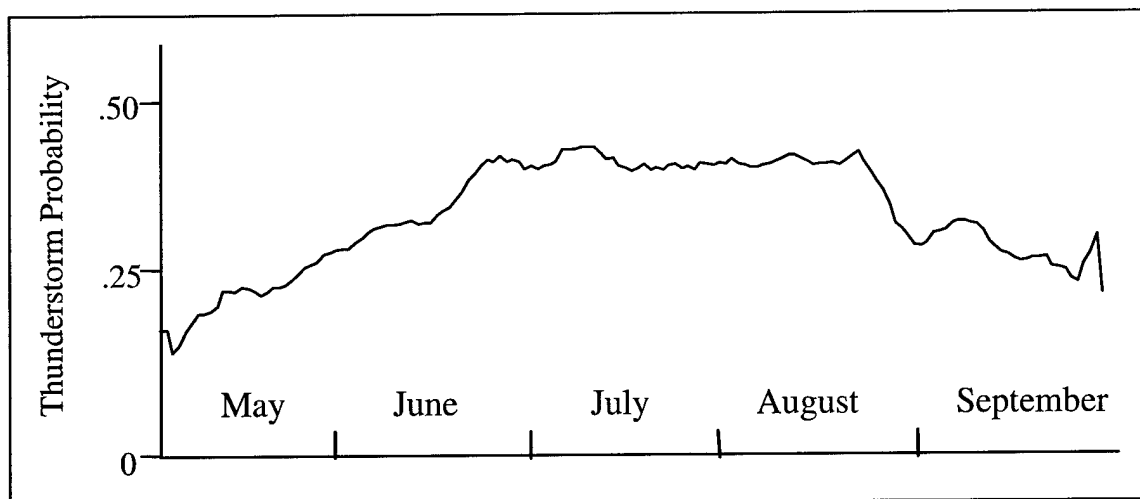
Figure 10 15-Day Weighted Moving Average of Climatological Probability
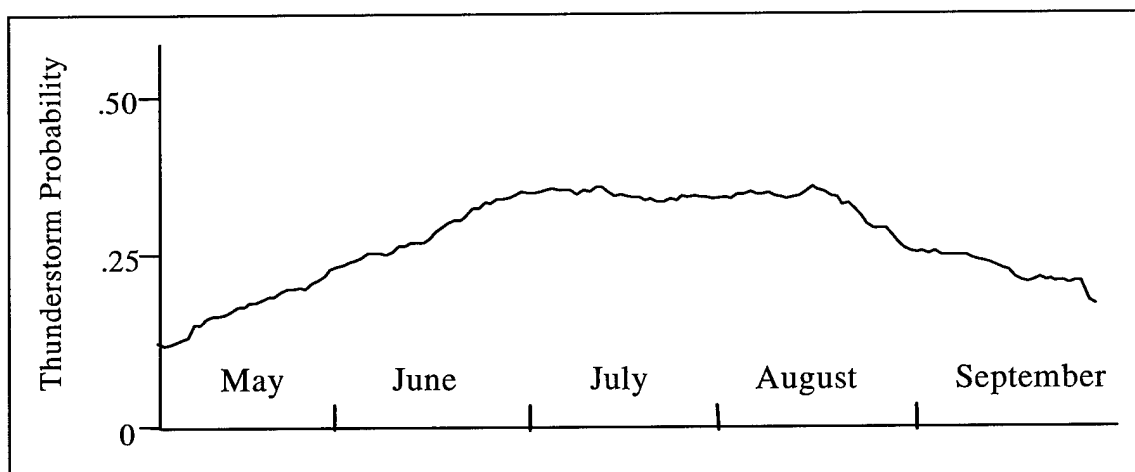
Figure 11 15-Day Linear Moving Average of Climatological Probability

To account for the fact that data is not available to be averaged within the first and last 7

days of the year, equation 29 was altered such that the summations used only 3-days.

Figure 10 illustrates the increased resolution available using the weighted average

method. Of particular interest were the twin local maxima of probability apparent at the

end of June and the beginning of July. This trend was noted by Neumann during his

analysis, but the additional data used in this research has reduced the strength of the

maxima (Neumann, 1968). These maxima were apparent in 90% of the individual yearly

45

records, yet the trends were not captured using a linear moving average as displayed in Figure 11. Also, a local minimum in probability became apparent at the end of August. The climatological probabilities using the 15-day weighted moving average were substituted for the day number in the day number function used by Neumann.

3.2.3 Multi-day Persistence

The conditional climatology of the various patterns of persistence is calculated to allow the inclusion of persistence as a variable during regression. The patterns of persistence are a sequence of thunderstorm occurrence or nonoccurrence. A thunderstorm occurrence is recorded as a 1, while nonoccurrence is recorded as a 0. The pattern is read from left to right with the most recent occurrence on the left. For example, 011 would represent no thunderstorms occurring yesterday, but thunderstorms occurring on the two days prior. These sequences were recorded as both a binary and decimal value during the initial data screening for thunderstorms. See Appendix A for the Fortran source code for creating the historical sequence. Also the conditional probabilities were calculated without using the data saved for verification. The binary representation was used for easier comprehension. The decimal representation is used to calculate the conditional climatology of 2-, 3-, 4-, 5-, and 6-day persistence patterns for each of the five months, using the method described in section 2.3. A Mathcad® 7.0 program (Appendix G) performed the calculations. The resulting probabilities are used as a regression variable.

3.3   Statistical Data Reduction

Because of the large amount of information available, various statistical analyses were used to determine the best regression variables. The less important variables were not considered for regression, so as to reduce the number of variables to a manageable number. The following sub-sections describe the two types of analysis used.

3.3.1   Point Biserial Correlation Coefficient (R_{pb})

Correlation was used to describe the association between variables and the record of thunderstorm occurrence. The 10 variables with the highest correlation were considered for regression. Measuring the association between continuous variables and dichotomous variables requires a type of correlation called point biserial correlation (Gibbons, 1976). The point biserial correlation is a reduced form of the Pearson Product-moment Correlation. The point biserial correlation coefficient is given by:

$$R_{pb} = \sqrt{\frac{n_1 \cdot n_0}{n}} \; \frac{\overline{X}_1 - \overline{X}_0}{\sqrt{\sum (x - \overline{X})^2}} \quad , \qquad (28)$$

where n_1 = number of thunderstorm occurrences,
   n_0 = number of nonoccurrences,
   n = total number of observations,
   $\overline{X}_1$ = mean of the variables paired with occurrences,
   $\overline{X}_0$ = mean of the variables paired with nonoccurrences,
   $\overline{X}$ = mean of all the variables
   X = variable of interest.

Table 6 Ten Highest Biserial Correlations

Thunderstorm Indices		Winds		Humidities	
Thompson Index	.371	850-mb U-wind	-.307	700-mb RH	.307
K Index	.360	900-mb U-wind	-.303	650-mb RH	.303
SSI	-.291	800-mb U-wind	-.296	1000-mb to700-mb RH	.293
		950-mb U-wind	-.294		

This type of correlation was used on all the variables in the data set and on 6 polynomial variations of the variables (Appendix F). The highest correlations are shown in Table 6.

Correlation analysis was performed on the conditional climatology and the climatological frequencies to determine if either showed any potential as a predictive variable in a regression. Also, the correlation coefficient was calculated for simple persistence. The results shown in Table 7 show that the probability determined from 6 day persistence is more predictive than simple persistence. In addition, the correlation coefficient was calculated for day number as used by Neumann and for raw frequency (unaltered data). The results also show that day number has the lowest correlation to thunderstorm occurrence while the raw frequency has the highest when comparing climatological frequency variables. The low correlation of day number is to be expected because day number makes no allowances for variability in the observed frequency. Also, the raw frequency was expected to have a higher correlation than the moving averages. This is because the smoothing of the moving averages reduces the correlation.

Table 7 Climatological Correlation Results

Conditional Probability (example)		Correlation	Climatological Frequency	Correlation
Simple Persistence	(1)	.329	Day number	.231
2-day Persistence	(11)	.388	Raw Frequency	.270
3-day Persistence	(001)	.395	3-day Weighted Average	.252
4-day Persistence	(0110)	.400	15-day Weighted Average	.243
5-day Persistence	(11010)	.411	15-day Linear Average	.236
6-day Persistence	(010110)	.441		

To determine which moving average to use, all moving average methods were individually regressed. The results of each regression were used to make forecasts using the independent data set, and the forecasts were measured for accuracy. The regression showed that the 15-day weighted average had the highest accuracy.

The correlation of wind direction versus thunderstorm occurrence was used to determine which wind sectors would best stratify the data sets. A Mathcad® 7.0 program was used to record whether the wind occurred within a sector (Appendix H). Another Mathcad® 7.0 program (Appendix I) was used to find the correlation of wind occurrence within a sector versus thunderstorm occurrence. The directions bounding the sectors were altered by 10° increments until the highest correlation was found for each pressure level. Figure 12 shows the sector with the highest correlation for each month.
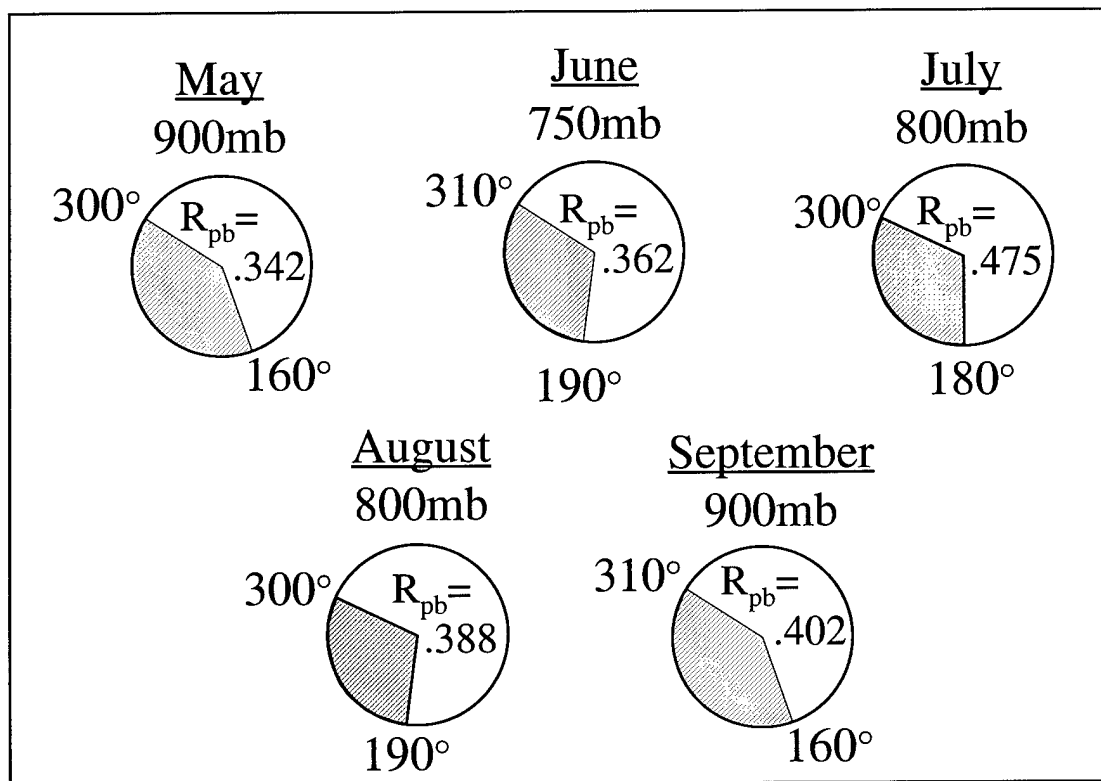
Figure 12 Highest Correlated Wind Sectors

49

3.3.2   Principal Component Analysis

Principal component analysis (PCA) was used to reduce the number of variables without losing too much of the original predictive information within the data set. The smaller number of variables used in linear combination still retain most of the variance contained within the original data set (Dillon and Goldstein, 1984). The PCA was run using S-plus® 4.5. The principal components were calculated using a correlation matrix as opposed to a covariance matrix because of the differing units of each variable. Using the correlation matrix prevents arbitrary magnitudes of variance from being created by the arbitrary scaling of the units (Wilks, 1995). The variables were divided into four groups and a PCA was performed on each group. The groups included thunderstorm indices, temperature- based variables, moisture-based variables, and wind-based variables. The 4 variables with the highest coefficients from each group were then used in a final PCA. The following variables from each group were identified as being able to explain the most variance when used in combination: Thompson index, 800-600-mb mean relative humidity, 550-mb heights, and the 750-mb U-component of the wind. These variables correspond well to the parameters from the literature review thought to cause thunderstorms.

After performing the correlation analysis and the principal component analysis, 8 variables were chosen for regression. These variables were the 850-mb winds, the 700-mb winds, the 600-mb winds, the Thompson index, the K index, the 800-600-mb mean relative humidity, the 6-day conditional climatology, and the 15-day weighted moving average. The algorithms produced have different combinations of these variables.

50

3.4 Algorithm Development

This section explains how both algorithms are created. The derivation for each is similar, even though the variables used differ. Even the creation of the categorically stratified algorithm uses the same steps, although it requires many more repetitions.

3.4.1 Logistic NPTI (LNPTI)

A logistic regression of the original NPTI variables was performed to judge their performance using a regression method other than linear regression. As in the original NPTI, each month was regressed separately. The regression was run using S-plus® 4.5 and all of the coefficients were gathered into one file. Finally, a prediction equation for each month was created by combining the coefficients with their respective functions.

3.4.2 Stratified Logistic Thunderstorm Index (SLTI)

The stratified logistic regression method is an extension of the idea of creating a regression equation for each month. This method is used to better represent the relationship between the wind variables and the predictand. This relationship changes statistically with each wind sector as shown in section 3.3.1. With stratification, the variables represented in each sector have to be regressed separately. The stratification method divides the data into months and then further divides each month's data set into wind sector data sets. These final data sets are the data sets regressed. For example, a regression is performed on variables observed on days in May with the winds at 900-mb from the sector 300° to 160°. A separate regression is performed on the variables observed in May with the winds at 900-mb outside the sector 300° to 160°.

51

To keep the number of variables for the final algorithm as small as possible, only the variables chosen in the previous section were used. This is necessary because stratification causes the number of coefficients needed to become prohibitively large. The chosen variables were the 850-mb winds, the 700-mb winds, the 600-mb winds, the Thompson index, the K index, the 800-mb to 600-mb mean relative humidity, the 6-day conditional climatology, and the 15-day weighted moving average. The coefficients for the Thompson Index for the Southeast wind sector during July were found to produce forecasts with worse accuracy than persistence. Fortunately, the coefficients for the Northeast sector produced forecasts with better accuracy than persistence even when used on variables from the Southeast sector. Therefore, July uses only Northeast sector coefficients for the Thompson index. To forecast using the resulting 545 coefficients, the day being forecast for must first be fit into a month and wind sector category. The month- and wind-dependent coefficients are then used with the 8 variables in the 2 levels of regression equations.

3.5   Verification

Each forecast method was verified to demonstrate its value and determine its effectiveness. The ten percent of data saved before regression began was used to complete the verification. Each forecast method was used to calculate and record the probability of a thunderstorm occurrence using the verification data set. Whether or not a thunderstorm occurred was also noted. This data was then used to estimate the accuracy of all four methods. Contingency tables of each of the possible cutoff values were created. Also, measures of accuracy and skill scores were calculated for each of the

entire range of possible cutoff values. Graphs, shown in chapter 4, display the results

created from the range of cutoff values. These graphs visually provide the best cutoff

value for a given accuracy measure. By displaying the results of all the methods on a

single graph, a comparison of each method's overall ability can be made. Finally,

persistence was assessed in the same manner.

The values comprising any specific contingency table can be determined from the

graph of all possible contingency tables. The lines labeled A through D inside Figure 13

represent the possible values for each of the similarly labeled four blocks in a

contingency table for an arbitrary forecast method. The abcissa represents the cutoff

values for each contingency table, and the ordinate is the number of forecasts for a

specified block of the contingency table at a given cutoff value. Each horizontal line

represents the values of a block in the contingency table for persistence and shows
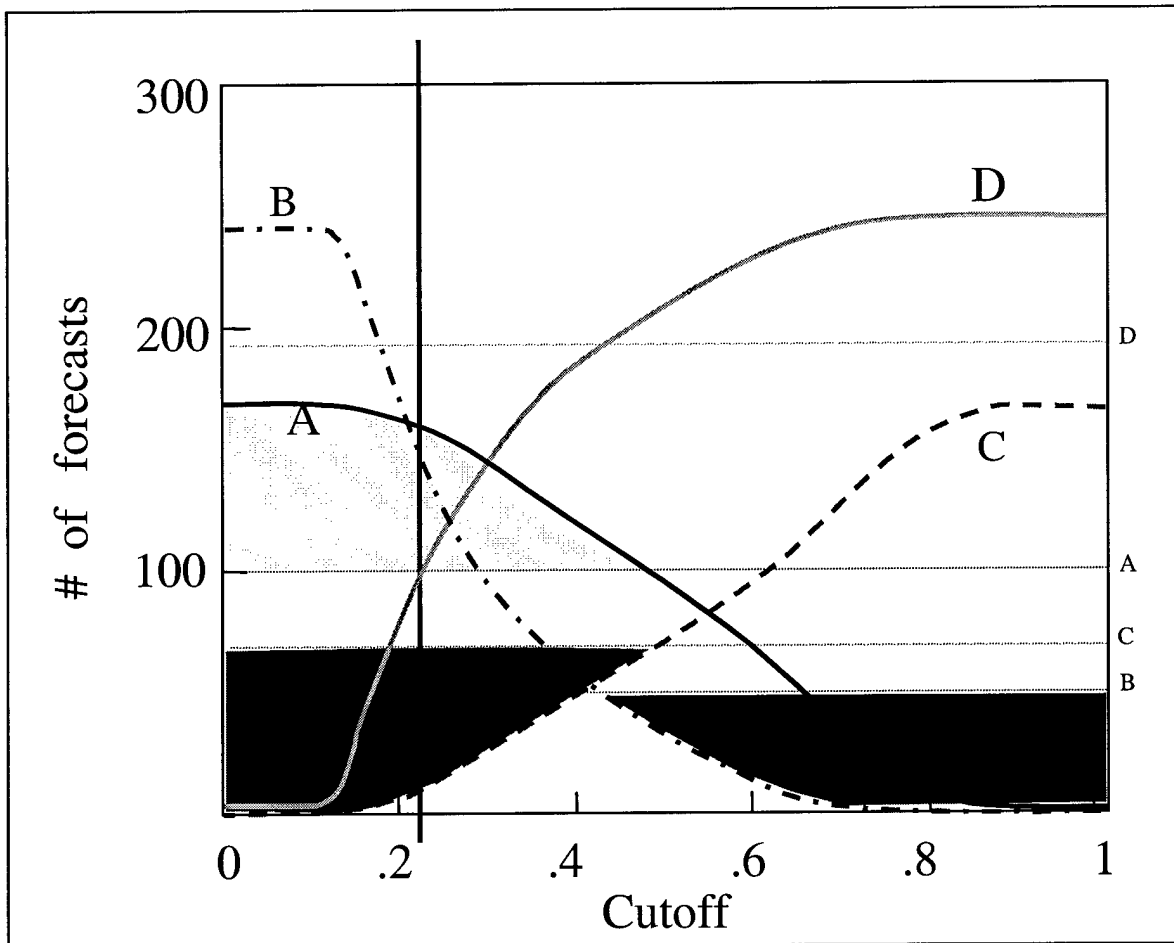
Figure 13 Hypothetical Example of Possible Contingency Table Graph

the number in each block remains constant for persistence. The other lines represent

blocks for the possible contingency tables for the arbitrary forecast method. A line drawn

vertically from a given cutoff value intersects each of the other lines. The ordinate value

at the intersection is the value for the representative block in a contingency table having a

cutoff equal to that of the ordinate.

Cutoff = .21						

Other	Observed		Persistence	Observed	
	Yes	No		Yes	No
Forecast Yes	156	147	Forecast Yes	100	50
Forecast No	13	97	Forecast No	69	194

Figure 14 Hypothetical Contingency Table

A vertical line at a cutoff of twenty-one percent would give the values shown in the contingency tables in Figure 14.

Only limited information can be gained about the accuracy of a forecast method by comparing corresponding blocks in two different contingency tables. Once a cutoff is chosen for a contingency table, no information is available about contingency tables with other cutoffs. The evaluation of one forecast method over another is better accomplished using a graph displaying all possible cutoffs for a contingency table. The accuracy of one method over another is then represented by the area between lines representing corresponding blocks. For lines A and D, areas above persistence are good, and for lines B and C, areas below persistence are good. The larger the area, the better one method is over the other. Graphs can also be made in a similar manner for the measures of accuracy. The accuracy of one method over another is then represented by the area between lines representing corresponding measures of accuracy.

4.   Results and Analysis

The NPTI was designed to predict thunderstorms using morning soundings. It was created at a time when the investigation of sea breeze mechanics, which play a large part in thunderstorm formation, had just begun. Also, the computer capacity to perform the necessary calculations was quite small. Since then, a better understanding of thunderstorm formation has been gained and computer power has increased dramatically. By using these advantages, an improvement in forecasting skill was realized. The results discussed in this section show the increase in accuracy obtained to date.

4.1   Persistence Results

Persistence is often used as a judge of other forecasting techniques. Therefore, persistence's forecast ability was evaluated and used as a milestone. No cutoff value is necessary because persistence is already a dichotomous value. The contingency table in Figure 15 contains the verification results from the independent data set. It is easy to see persistence forecast correctly more often than it forecast incorrectly. Of the 413 forecasts verified, 294 were forecast correctly and 119 were forecast incorrectly.

		Observed	
		Yes	No
Forecast	Yes	A 100	B 50
	No	69 C	194 D

Figure 15 Contingency Table for Persistence Forecast

56

	BS	HR	TS	TSN	POD	PODN	FAR	FARN
Persistence	28.8	71.2	46.0	62.0	59.2	79.5	33.3	26.2

Table 8 Accuracy Measures (%) for Persistence

Table 8 lists the various skill scores for persistence. The threat scores (TS and TSN) and

probabilities of detection (POD and PODN) show that correctly forecasting "no

thunderstorm" is the reason that hit rate (HR) is high. The False Alarm Rate (FAR) is

33.3%, while the value block c in the contingecy table indicates persistence

underforecasts thunderstorms. Despite these problems, persistence does well with a Brier

Score (BS) of 28.8%.

4.2   NPTI Results

The performance of NPTI was evaluated using the statistical methods described in

chapters two and three. The Neumann-Pfeffer Thunderstorm Index did poorly compared

to persistence. Figure 16 shows that no cutoff can be chosen so that all four blocks of the

NPTI contingency table are simultaneously better than persistence. One of the better

cutoffs found was 21%. The contingency tables for this cutoff are shown in Figure 17

and the accuracy measures using 21% are shown in Table 9. In all cases except POD,

Table 9 shows NPTI does worse than persistence. This is more easily seen by looking at

Figure 18. The HR and TS never do better than persistence. TS and POD do increase as

the cutoff is lowered, but the value in block b in Figure 16 shows this gain is offset by an

increase in false alarms. In addition, the skill scores showed poor performance compared

57
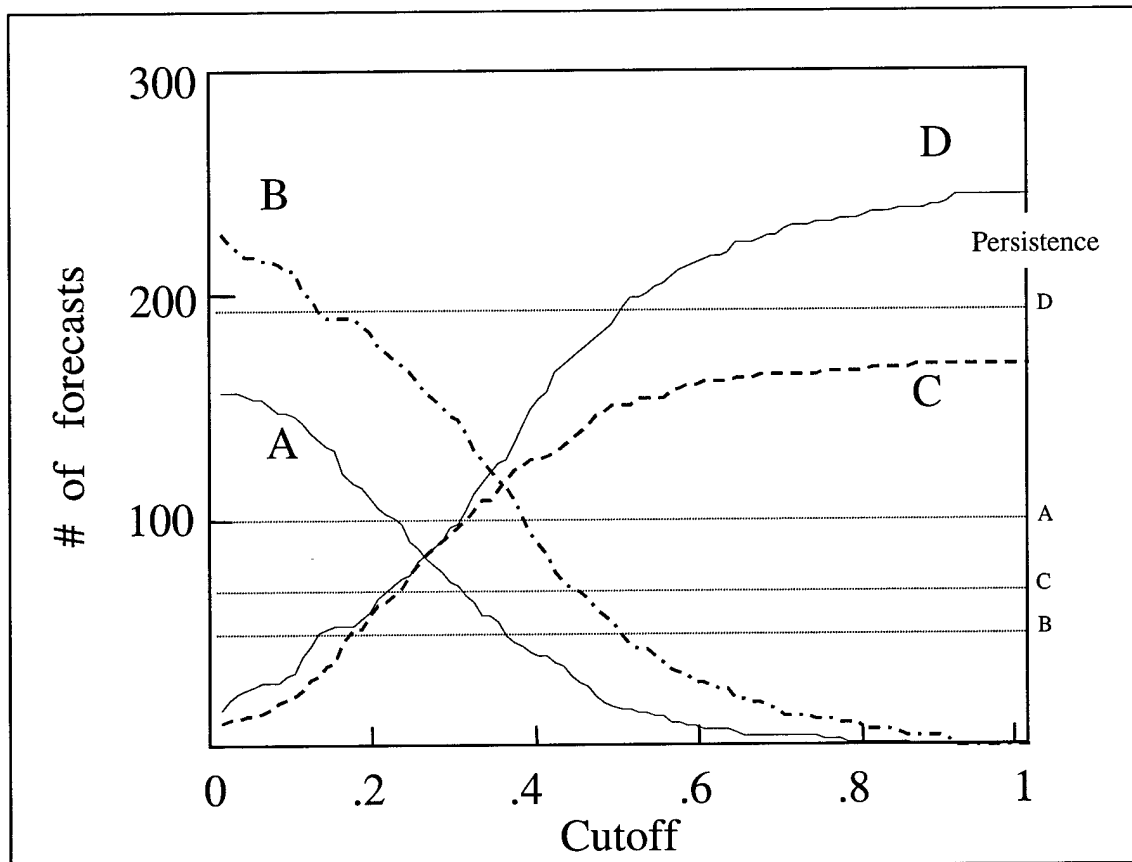
Figure 16 Graph of Possible Contingency Tables of NPTI

to persistence. A χ^2 of 4.723 at 95% confidence level showed a small dependence

between forecast and observation. Therefore, NPTI's accuracy may be random chance.

Figure 17 Contingency Tables for NPTI (Cutoff .21) and Persistence

Table 9 NPTI Accuracy Measures (Cutoff .21)

	BS	HR	TSY	TSN	POD	PODN	FAR	FARN	HSS	KSS
NPTI	.320	.420	.302	.233	.615	.283	.627	.485	-.093	-.10
Persistence	.288	.712	.460	.620	.592	.795	.333	.262	.394	.387

Skill Scores (%)	RSS	SS HR	SS TS	SS TSN	SS POD	SS PODN	SS FAR	SS FARN
	-11.1	-101	-28.4	-104	5.8	-250	-88.2	-84.9
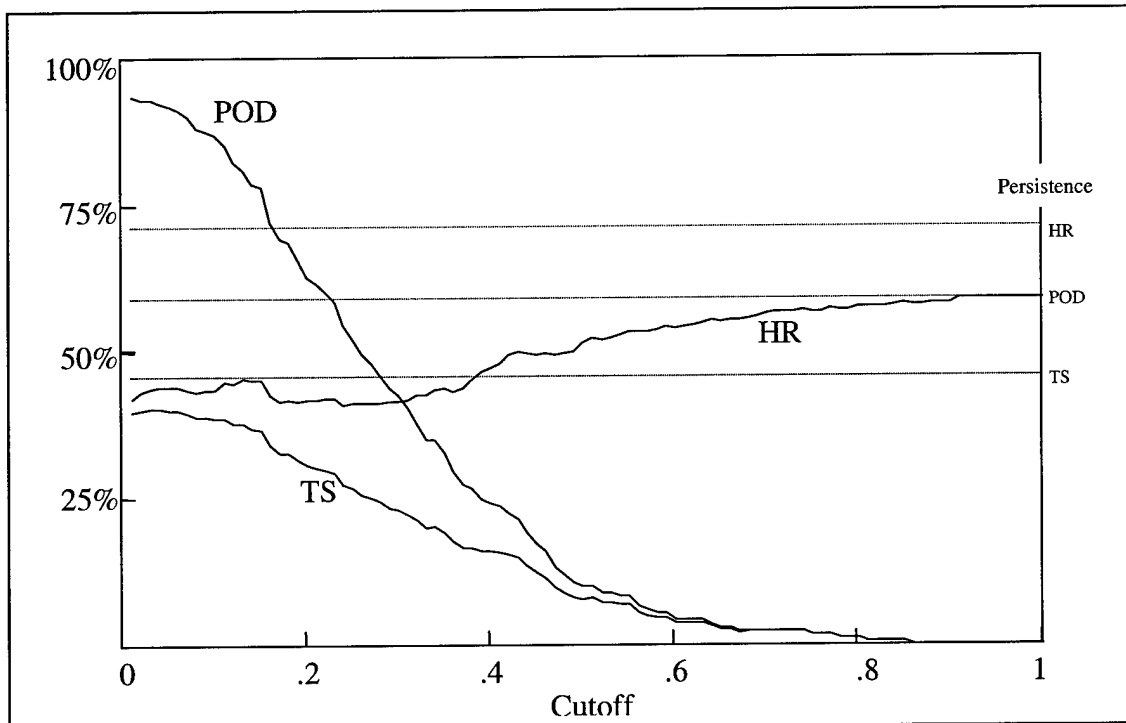
Figure 18 Graph of Possible NPTI Accuracy Measures

4.3   New Algorithm Results

The newly derived algorithms did much better than NPTI when compared to persistence. Without choosing a cutoff, all three algorithms achieved a better Brier Score. Table 10 shows that SLTI has a 47% better BS than persistence, and Table 11 shows that LNPTI has a 44% better BS than persistence. Because persistence had a better BS than NPTI, the two new methods were also better than NPTI. The ratio skill score, using NPTI as the reference, was 49.75%Choosing a cutoff that caused all four blocks of the contingency table to be simultaneously better than persistence or NPTI was possible using either new method. The highest skill scores were found when using a cutoff close to the cutoff where lines B and C intersect in Figure 19 and in Figure 22. A cutoff of 44% was used to create the contingency tables in Figure 20 and in Figure 23 and gives a HR of 77.2% for SLTI and 75.1% for LNPTI.

Both SLTI and LNPTI displayed the large ranges of cutoff values where the forecast method performed better than persistence. These areas are shaded in Figure 21. Any choice of cutoff between 20% and 60% for SLTI provides a higher hit rate, threat score, or probability of detection than persistence. For LNPTI, Figure 24 shows that any choice of cutoff between 25% and 55% provides a higher hit rate, threat score, or probability of detection than persistence. All skill scores for SLTI in Table 10 show at least 12% better than persistence. A χ^2 for SLTI of 128.5 at 95% confidence level showed a strong dependence between the forecast and observation. SLTI had the best performance overall.
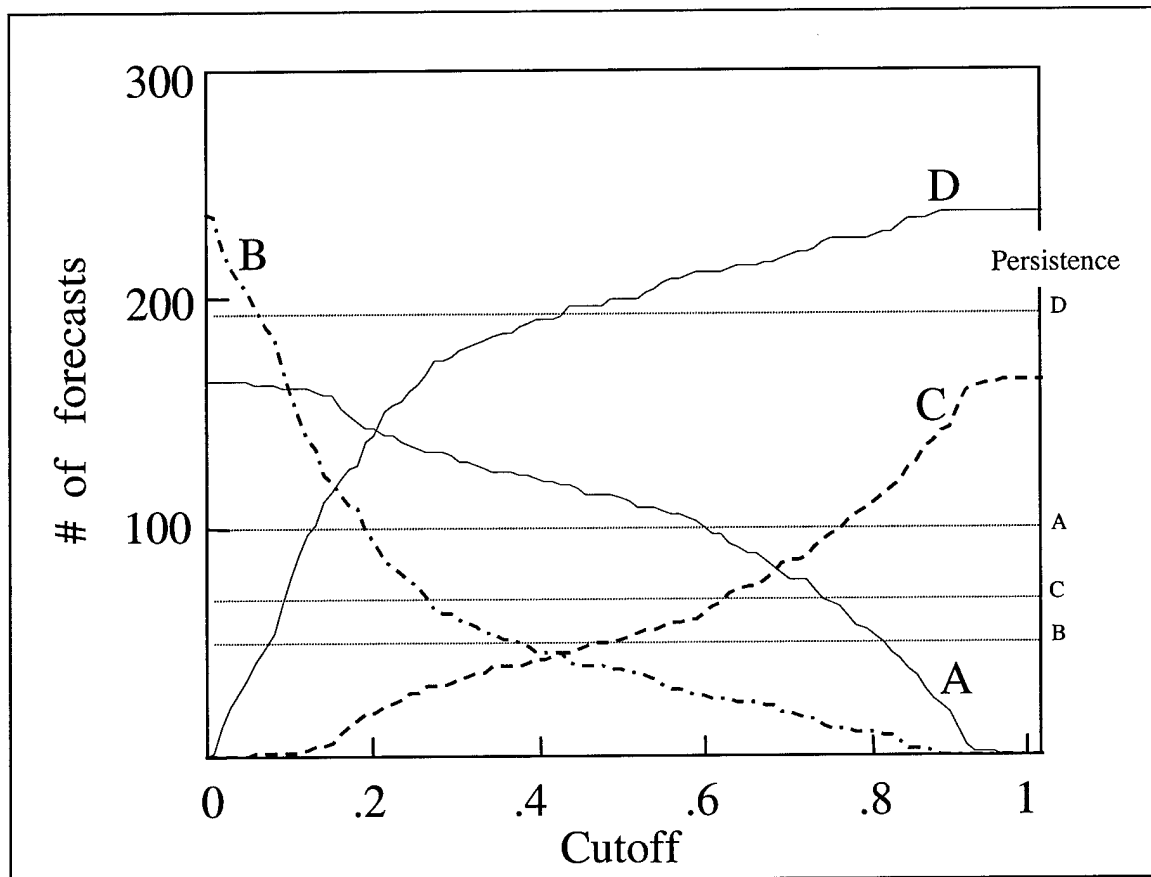
Figure 19 Possible Contingency Tables for SLTI

Figure 25 through Figure 27 show how much better SLTI was than NPTI. The distance between corresponding lines is a measure of improvement of one method over the other. The graphs show large areas between the lines that indicate SLTI is better than LNPTI.

Cutoff = .44

	Observed				Observed	
SLTI	Yes	No		Persistence	Yes	No
Yes	A 119	44 B		Yes	A 100	50 B
Forecast				Forecast		
No	50	200		No	69	194
	C	D			C	D

Figure 20 Contingency Tables for SLTI (Cutoff .44) and Persistence
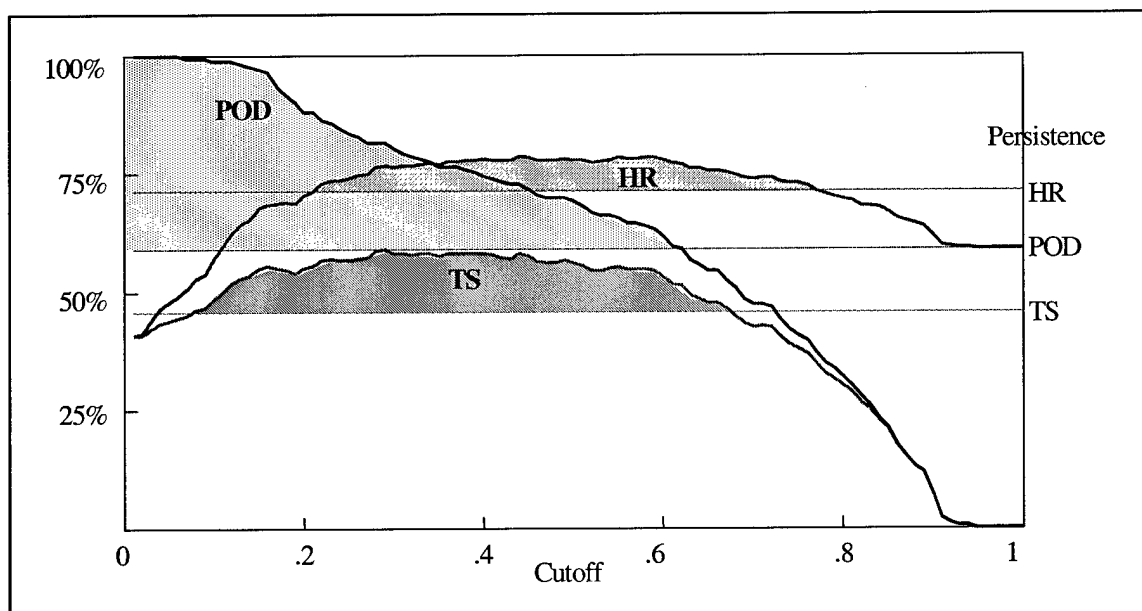
Figure 21 All Possible SLTI Accuracy Measures

Table 10 SLTI Accuracy Measures (Cutoff .44)

	BS	HR	TS	TSN	POD	PODN	FAR	FARN	HSS	KSS
SLTI	.161	.772	.558	.680	.704	.829	.270	.200	.527	.524
Persistence	.288	.712	.460	.620	.592	.795	.333	.262	.262	.890

Skill Scores (%)	RSS	SS HR	SS TS	SS TSN	SS POD	SS PODN	SS FAR	SS FARN
	44.2	21	18.8	15.9	27.5	12	19	23.8
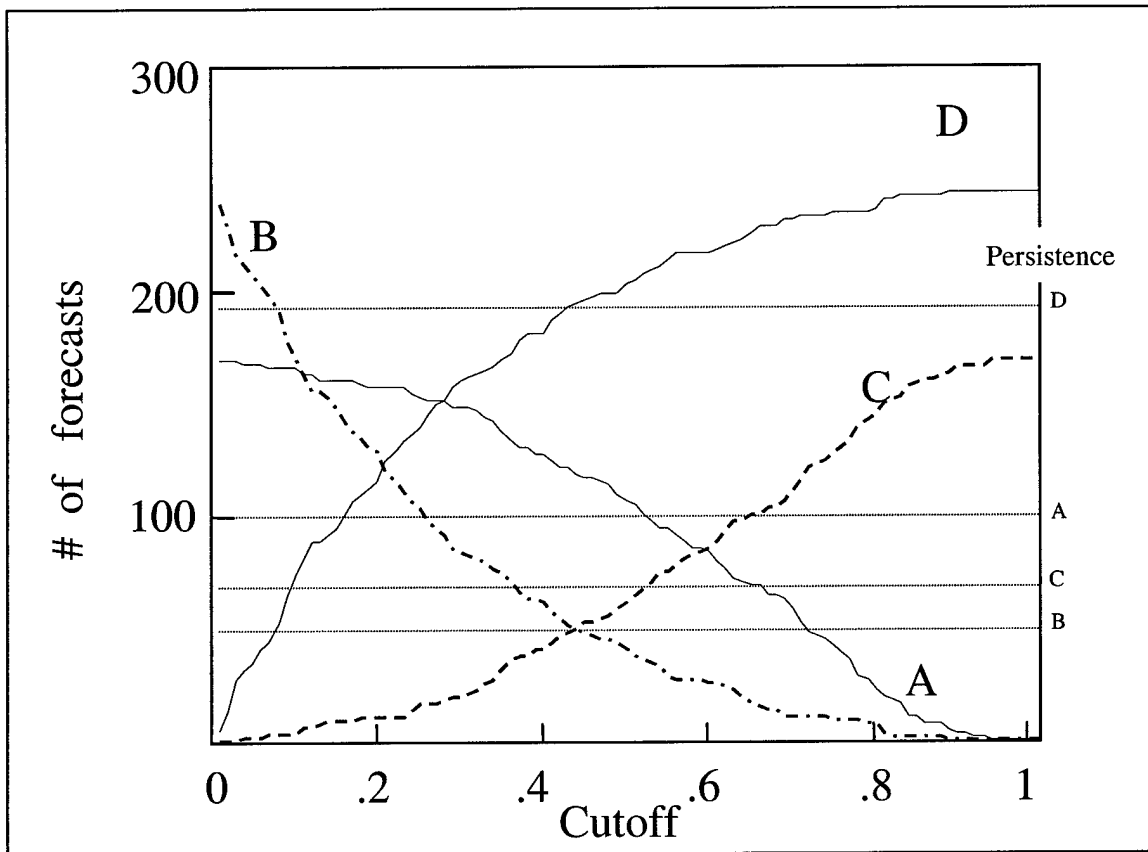
Figure 22 Possible Contingency Tables for LNPTI

Cutoff = .44

	Observed	
LNPTI	Yes	No
Forecast Yes	A 119	B 49
Forecast No	C 50	D 195

	Observed	
Persistence	Yes	No
Forecast Yes	A 100	B 50
Forecast No	C 69	D 194

Figure 23 Contingency Tables for LNPTI (Cutoff .44) and Persistence
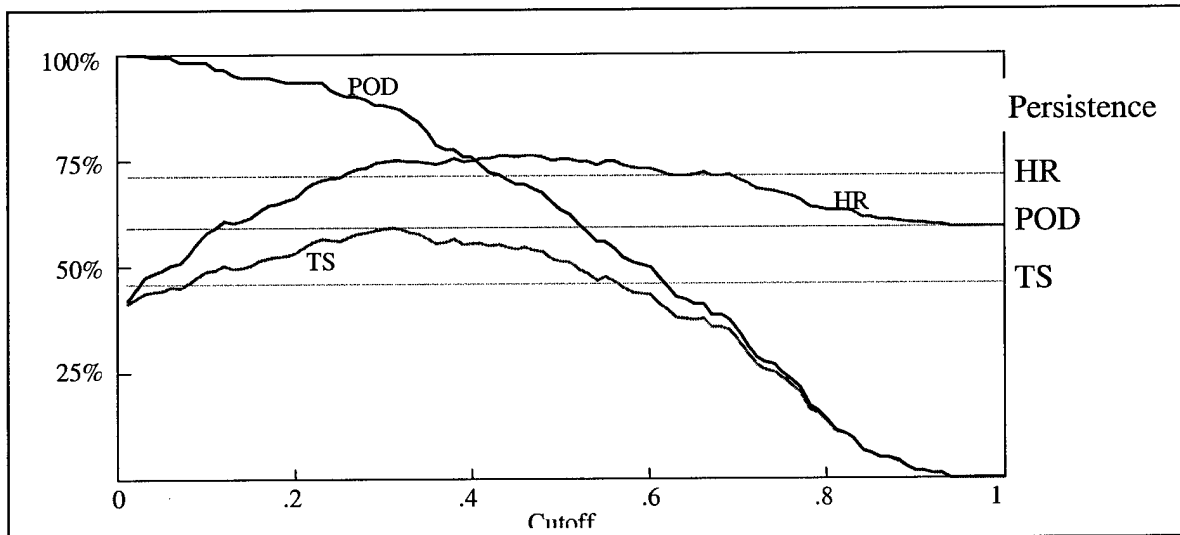
Figure 24 LNPTI Accuracy Measures Using All Possible Cutoff Values

Table 11 LNPTI Accuracy Measures (Cutoff .44)

	BS	HR	TSY	TSN	POD	PODN	FAR	FARN	HSS	KSS
LNPTI	.165	.76	.546	.663	.704	.80	.29	.204	.504	.503
Persistence	.288	.712	.460	.620	.592	.795	.333	.262	.262	.890

Skill Scores (%)	RSS	SS HR	SS TS	SS TSN	SS POD	SS PODN	SS FAR	SS FARN
	42.8	16.8	16.4	11.4	27.5	2	12.5	22.2
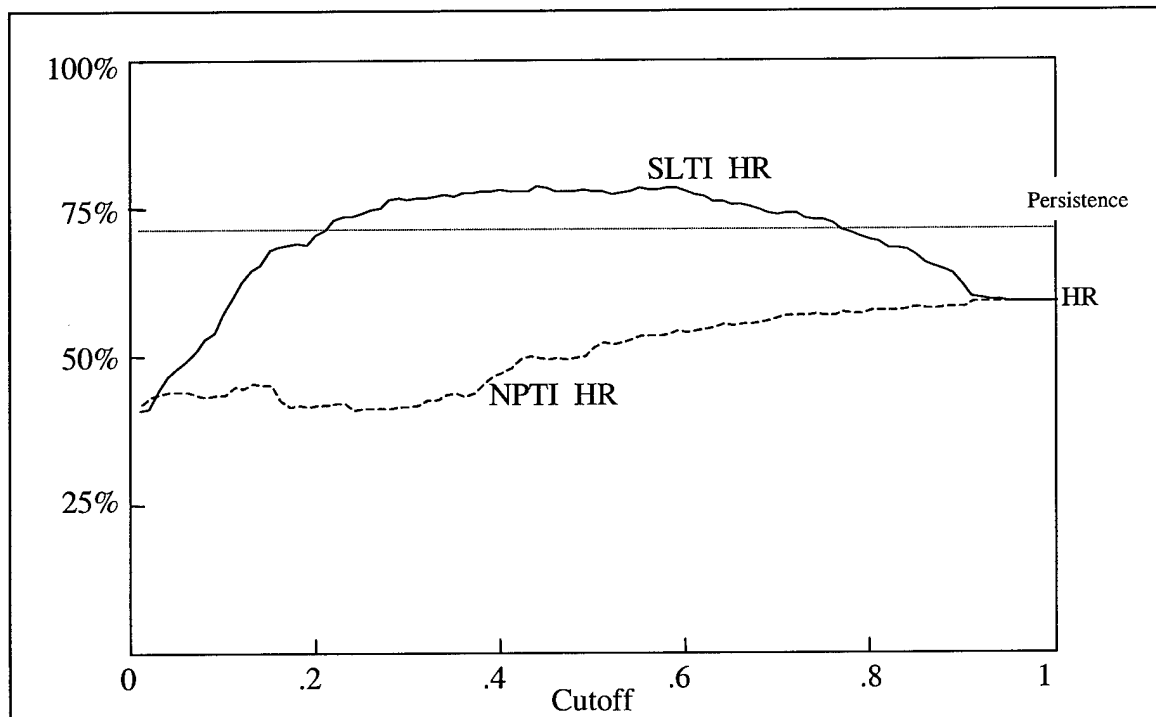
Figure 25 SLTI and NPTI Hit Rate
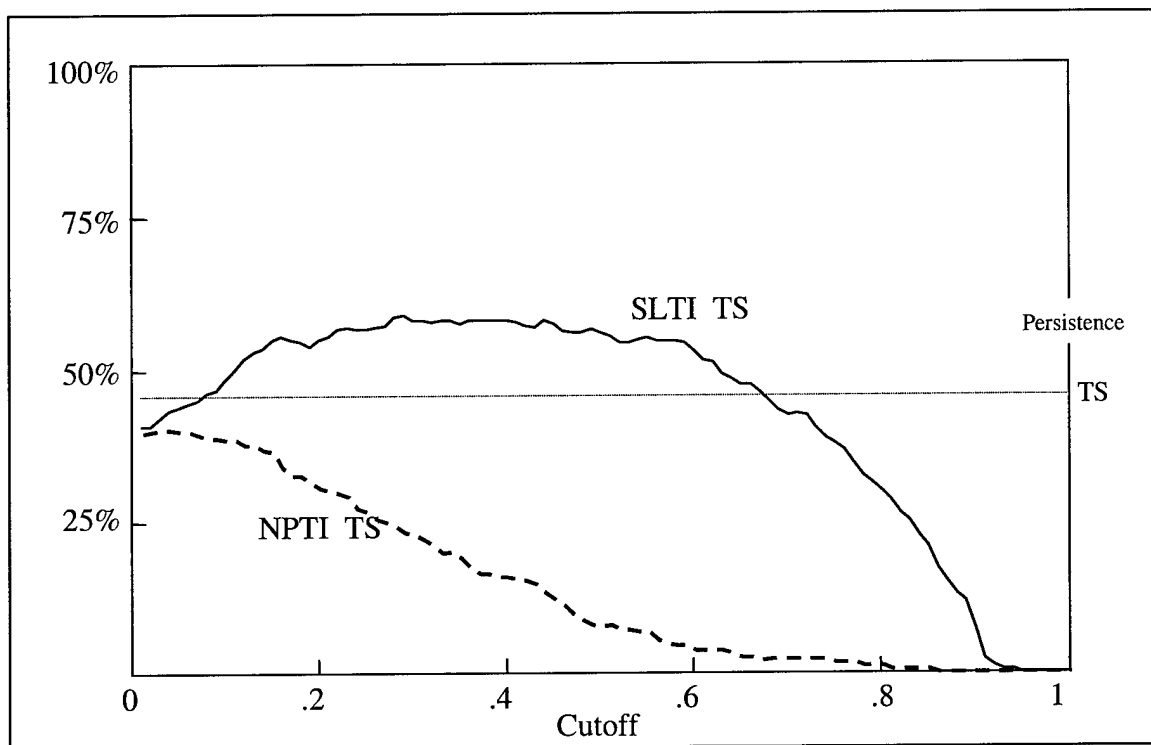
Figure 26 SLTI and NPTI Threat Score
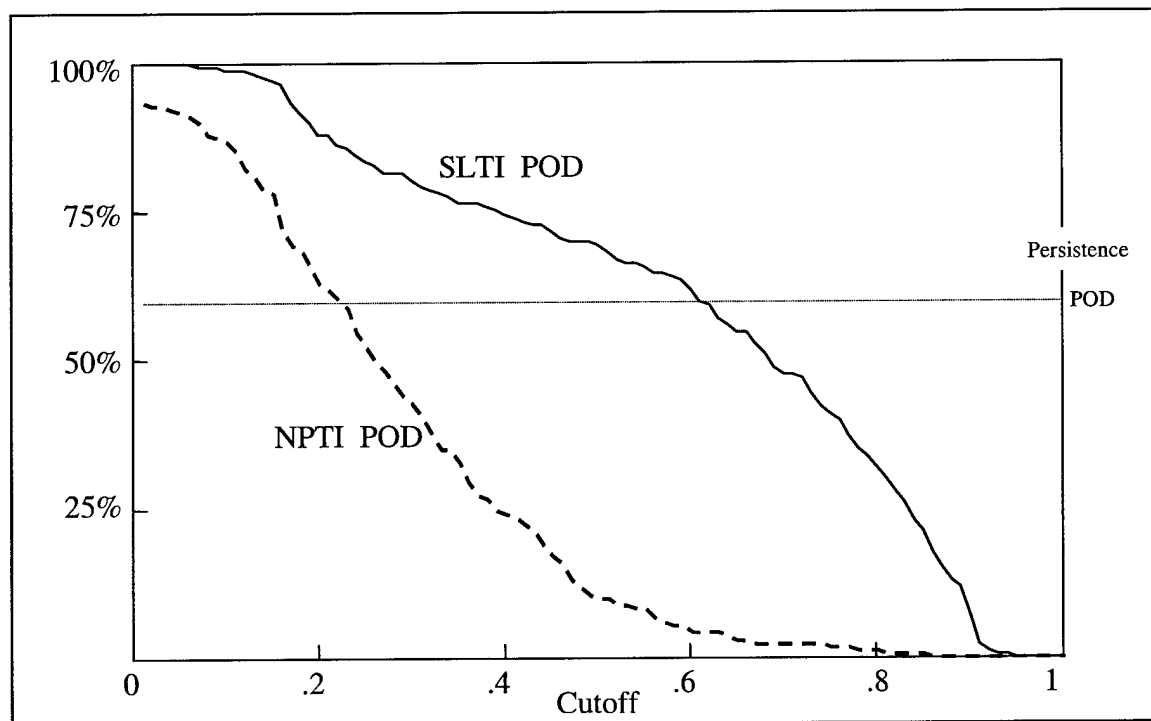
Figure 27 SLTI and NPTI Probability of Detection

5.   *Conclusions and Recommendations*

This study began as an attempt to verify the idea that logistic regression could improve the forecast ability of the NPTI algorithm. The research revealed that using other variables might also provide improved capability. The goal then expanded to include the search for the variables which could provide the highest increase in accuracy. Because all possible combinations of variables were not tried, further increases in accuracy may be possible by finding better variables to regress.

5.1   Conclusions

Logistic regression creates a more accurate forecast algorithm than linear regression when forecasting thunderstorms at Cape Canaveral. By using the same variables and only changing the regression method, LNPTI shows significant improvement over NPTI. LNPTI's highest hit rate is 17 percent higher than the highest hit rate NPTI was able to achieve. Also, there is more than one cutoff that allows the values in every block in the LNPTI contingency table to be better than the corresponding NPTI contingency table values.

Two additional changes to NPTI show a further increase in forecast accuracy. The first change, using separate regressions for each wind sector during each month, allow the accuracy of each regression to be maximized for the variables used. By using separate regressions for each sector, the regression coefficients are tuned to the two

synoptic regimes that most influence thunderstorm occurrence. The different regressions create coefficients that best forecast thunderstorms given the different environments of the regimes. The second change, altering the variables used for regression, increases the predictive capability of the algorithms. One factor causing an increase in forecast capability is the change in the thunderstorm index used as a variable. Both the Thompson index and K index shows higher correlation to thunderstorm occurrence than the SSI, and test regressions of each of the variables individually indicate SSI performs the worst of the three. Therefore, replacing SSI with the Thompson index and the K index increases the accuracy of the algorithm. Additionally, regressing multiple wind levels that more closely correspond to those seen during sea breezes gives the regression more discriminatory power. These levels also correspond to those mentioned in the literature review and correlation analysis. Each of these changes in variables adds a little more to the predictive power of the algorithm.

5.2   Recommendations

NPTI should be replaced with an algorithm created by logistic regression. Logistic regression has been shown to improve the performance of the methods created by Neumann and to have more accuracy than persistence. All three new methods show similar forecast ability, but the Stratified Logistic Thunderstorm Index has the highest hit rate, threat score, and Brier score. The accuracy of NPTI has been demonstrated to be very poor in this study. NPTI's accuracy is even worse than persistence, and this finding agrees with previous studies. Because of NPTI's poor performance, the Stratified Logistic Thunderstorm Index should be implemented as soon as possible.

5.3 Suggestions for Future Research

Future study of this subject could go in two general directions. The method of regression used with NPTI can be further researched or new methods can be tried. Continued experimentation with combinations of variables may produce further improvement in accuracy. Testing the components of the thunderstorm indices with a decomposition of the variables could provide insight into which environmental parameters would work the best. Different variables in combination with further categorization of synoptic regimes for each regression may also provide an increase in forecast accuracy. Further research of the affect of dividing regressions by wind sector and month would reveal how much accuracy can be improved in this manner.

Two entirely different approaches are also likely to yield an increase in accuracy. The first is to use the wind tower data to better take advantage of the predictive information in the sea breezes. One morning sounding is not capable of providing the resolution necessary for understanding the complex pattern of sea breezes in the Cape Canaveral area. Since sea breezes play such an important part in thunderstorm production, more detailed information on their activity should be used in forecasting. The numerous wind towers in the Cape Canaveral area would give more information on the creation and strength of the local sea breezes. Another approach would be to use multi-station regression instead of single station regression. Single station regression has little prediction capability for air masses that advect into the area after the morning sounding. If the environment changes the probability of thunderstorms will change. Having more than one station included in the regression will allow changes in the

environment to be detected before they reach Cape Canaveral. Either approach could

provide a better forecast for the 45th Weather Squadron.

APPENDIX A. FORTRAN CODE FOR SCREENING
SURFACE OBSERVATIONS

```
PROGRAM tcount                    !for 1950-1996
!**********************************************************************
!This program converts observations to thunderstorm days.  A thunderstorm day is defined
!as TS,TSRA,+TSRA,or -TSRA occurring at least once per local day.  Because
!observations are recorded in UTC time but thunderstorm days are defined in local time,
!date change must be considered. 1100UTC through 2400UTC observations are counted toward the
!current day.  0000UTC through 0400UTC observations are counted toward the previous day.
!**********************************************************************
    IMPLICIT NONE
    INTEGER, Parameter :: x=5              !creates binary history of x # of days
    INTEGER :: YR,MON,DAY,HR,MIN,F,G,H,I,J,K,L,m,N
    INTEGER, DIMENSION(47,5,31) :: COUNT
    CHARACTER, DIMENSION(x) :: D*1
    CHARACTER :: ID*4,CODE*5,COMMA*1,WX*5,REST*5,DIR*3,SPD*2,VIS*6

    OPEN (45,FILE='970847.txt',STATUS='OLD')
    OPEN(21,FILE='frequency5.txt')

    L=0
    count=0
    DO WHILE (L.NE.1)
    READ (45,*,END=999) YR,MON,DAY,ID,CODE,HR,MIN,DIR,SPD,COMMA,VIS,WX,REST

    IF ((MON.gt.4).and.(MON.lt.10)) THEN
!**********************************************************************
!Counts Thunderstorms occurring today (UTC and local)
!**********************************************************************
     IF (HR.gt.10) THEN
       IF ((WX.EQ."TS").or.(WX.EQ."TSRA").or.(WX.EQ."+TSRA").or.(WX.EQ."-TSRA")) THEN
         COUNT(YR-49,MON-4,DAY)=1
       ENDIF
     ENDIF
!**********************************************************************
!Counts Thunderstorms occurring yesterday local and today UTC
!**********************************************************************
     IF (HR.lt.5) THEN
       IF ((WX.EQ."TS").or.(WX.EQ."TSRA").or.(WX.EQ."+TSRA").or.(WX.EQ."-TSRA")) THEN

         !**************************************************
         !Counts thunderstorms occurring on June 31st local
         !and July 1st UTC
         !**************************************************
         IF ((MON.eq.7).and.(DAY.eq.1)) COUNT(YR-49,2,30)=1

         !**************************************************
         !Counts thunderstorms occurring on the last day of
         !the month, except for June
         !**************************************************
         IF ((MON.ne.7).and.(DAY.eq.1)) COUNT(YR-49,MON-5,31)=1
```

```
!**************************************************
!Counts thunderstorms occurring on any day except
!the last day of the month
!*************************************************
     IF (DAY.ne.1) COUNT(YR-49,MON-4,DAY-1)=1
   ENDIF
  ENDIF
ENDIF


!*******************************************************************
!Counts Thunderstorms occurring Sept 30th local and Oct 1st UTC
!*******************************************************************
   IF ((MON.eq.10).and.(DAY.eq.1).and.(HR.lt.5)) COUNT(YR-49,5,30)=1


   L=0
   WX=" "
   ENDDO
999  L=0



!*******************************************************************
!Output to a file
!*******************************************************************
     H=0                    !Whole day number
     DO K=1,47
       G=0                  !Year day number
       F=0                  !Persistence
       D='0'                !History in binary
       DO J=1,5
         DO I=1,30
           H=H+1
           G=G+1
           N=0              !History in decimal

         DO m=1,x                !converts binary history to decimal
           IF (D(m).eq.'1') N=N+2**(x-m)
         ENDDO

         WRITE(21,*) COUNT(K,J,I),' ',H,' ',G,' ',F,' ',(D(m),m=1,x),' ',N
         F=COUNT(K,J,I)

         DO m=x,2,-1                !converts binary history to decimal
           D(m)=D(m-1)
         ENDDO

         IF (F.eq.1) THEN        !records persistence into history in binary
           D(1)='1'
         Else
           D(1)='0'
         ENDIF

         IF ((J.ne.2).and.(J.ne.5).and.(I.eq.30)) THEN
           H=H+1
           G=G+1
           N=0
```

72

```
      DO m=1,x                  !converts binary history to decimal
        IF (D(m).eq.'1') N=N+2**(x-m)
      ENDDO

      WRITE(21,*) COUNT(K,J,31),' ',H,' ',G,' ',F,' ',(D(m),m=1,x),' ',N
      F=COUNT(K,J,31)

      DO m=x,2,-1               !converts binary history to decimal
        D(m)=D(m-1)
      ENDDO

      IF (F.eq.1) THEN          !records persistence into history in binary
        D(1)='1'
      Else
        D(1)='0'
      ENDIF

    ENDIF
   ENDDO
  ENDDO
 ENDDO

25  FORMAT (31I2)
35  FORMAT (47I3)


END
```

APPENDIX B. FORTAN CODE FOR SCREENING
UPPER AIR OBSERVATIONS

```fortran
PROGRAM UAbyhour                              !for 1950-1996
!********************************************************************
!This program sorts Upper Air data. Only summer months with data between 09UTC to 16UTC are saved.
!********************************************************************
      INTEGER :: YR,MON,DAY,HR,PRESSURE,HGT,L,H
      CHARACTER :: MONTH*3,TEMP*3,DPT*3,DIR*3,SPD*3,RH*3
      OPEN (75,FILE='uadata.txt',STATUS='OLD')
      OPEN(19,FILE='UA9.txt')
      OPEN(20,FILE='UA10.txt')
      OPEN(21,FILE='UA11.txt')
      OPEN(22,FILE='UA12.txt')
      OPEN(23,FILE='UA13.txt')
      OPEN(24,FILE='UA14.txt')
      OPEN(25,FILE='UA15.txt')
      OPEN(26,FILE='UA16.txt')
      L=0
      DO WHILE (L.NE.1)
      READ (75,*,END=999) HR,DAY,MONTH,YR,PRESSURE,HGT,TEMP,DPT,DIR,SPD,RH
!********************************************************************
!Converts Text month to number
!********************************************************************
      IF (MONTH.eq.'MAY') MON=5
      IF (MONTH.eq.'JUN') MON=6
      IF (MONTH.eq.'JUL') MON=7
      IF (MONTH.eq.'AUG') MON=8
      IF (MONTH.eq.'SEP') MON=9
      IF ((MONTH.ne.'MAY').and.(MONTH.ne.'JUN').and.(MONTH.ne.'JUL').and.&
      &(MONTH.ne.'AUG').and.(MONTH.ne.'SEP')) THEN  MON=0
      ENDIF
!********************************************************************
!Allows only summer months between certain hours to be saved
!********************************************************************
     IF ((MON.ge.5).and.(MON.le.9)) THEN
       IF ((HR.ge.9).and.(HR.le.16)) THEN
         WRITE((HR+10),85) HR,DAY,MON,YR,PRESSURE,HGT,' ',TEMP,' ',DPT,' ',DIR,' ',SPD,' ',RH
       ENDIF
     ENDIF
     ENDDO
999 L=0
!********************************************************************
! This section puts in the end observations needed to use combine.f90
!********************************************************************
     DO H=9,16
       WRITE((H+10),87)  H,'1','1','1998',' 800',' 999',' 999',' 999',' 999',' 999',' 999'
       WRITE((H+10),87)  H,'1','1','1998',' 900',' 999',' 999',' 999',' 999',' 999',' 999'
       WRITE((H+10),87)  H,'1','1','1998','1000',' 999',' 999',' 999',' 999',' 999',' 999'
     ENDDO
85  FORMAT (I2,' ',I2,' ',I1,' ',I4,' ',I4,' ',I5,10a)
87  FORMAT (I2,' ',a1,' ',a1,' ',a4,' ',a4,6a4)
     END
```

APPENDIX C. INPUT CONSTANTS FOR CURRENT NPTI

F(X1) May	F(X1) June	F(X1) July	F(X1) August	F(X1) September
0.1727535E+00	0.3593742E+00	0.4379572E+00	0.4354426E+00	0.2868158E+00
0.1051405E-01	0.2334107E-01	0.4200581E-01	0.3692987E-01	0.1150187E-01
0.1604004E-01	0.1553199E-01	0.9613875E-02	-0.5944290E-03	0.7388294E-02
0.3773833E-03	0.1121490E-03	0.2973285E-03	-0.4935843E-05	0.1446295E-03
-0.1206595E-03	-0.5153501E-03	-0.1928031E-03	0.2941582E-03	-0.1896570E-03
0.2659499E-03	-0.1334638E-03	-0.1363037E-02	-0.6893766E-03	-0.9508320E-04
-0.1010322E-04	-0.1292631E-05	-0.5532370E-04	-0.6002453E-04	-0.9760508E-05
-0.2995797E-04	-0.1528670E-05	0.5288076E-04	-0.3215410E-05	-0.1847868E-05
0.2443706E-04	-0.2406814E-04	-0.4731448E-04	-0.1252746E-04	0.4754807E-05
-0.8702967E-05	-0.2729250E-06	0.2012023E-04	0.5158292E-04	-0.2164692E-05

F(X2) May	F(X2) June	F(X2) July	F(X2) August	F(X2) September
0.1222090E+00	0.3322431E+00	0.4401347E+00	0.4004517E+00	0.2534370E+00
0.8970767E-02	0.2193566E-01	0.3111313E-01	0.3162476E-01	0.9343341E-02
0.1238091E-01	0.1054781E-01	0.1254637E-02	0.2487462E-02	0.5741805E-02
0.1530813E-03	0.9557068E-04	0.1287248E-03	0.8781690E-04	0.6548849E-04
0.1435101E-04	-0.1835915E-03	0.1927784E-04	0.8207381E-05	-0.1114670E-03
0.2574404E-03	0.3771296E-04	-0.1803830E-03	0.2306297E-03	-0.5738156E-04
-0.3452270E-05	-0.1445187E-04	-0.2351527E-04	-0.3665728E-04	-0.3442634E-05
-0.1025880E-04	-0.9028485E-05	0.1468542E-04	-0.1493087E-04	-0.6619258E-05
-0.7730953E-05	0.2304699E-05	-0.5562572E-05	-0.1199524E-04	0.4766845E-05
-0.1276384E-05	-0.6225688E-05	-0.5263795E-05	-0.1924274E-05	0.5612928E-05

F(X3) May	F(X3) June	F(X3) July	F(X3) August	F(X3) September
0.6787956E-01	0.1357487E+00	0.3103540E-01	0.1167747E+01	0.7477421E-01
-0.7066473E-02	-0.1913794E-01	-0.1501027E-01	-0.7547790E-01	-0.1267558E-01
0.3165507E-03	0.7981991E-03	0.7151766E-03	0.1655678E-02	0.4923630E-03
-0.1972934E-05	-0.6220716E-05	-0.5382679E-05	-0.9955816E-05	-0.3633523E-05

F(X4) May	F(X4) June	F(X4) July	F(X4) August	F(X4) September
0.4313504E+00	0.5758111E+00	0.5912066E+00	0.5572201E+00	0.3793839E+00
-0.7493180E-01	-0.6988162E-01	-0.5364221E-01	-0.4718516E-01	-0.5686991E-01
0.2996801E-02	0.1929542E-02	0.5560704E-03	-0.1221637E-02	0.2028510E-02

F(X5) May	F(X5) June	F(X5) July	F(X5) August	F(X5) September
-0.4076566E+01	-0.9433578E+01	-0.1056197E+01	-0.1015726E+02	0.1715063E+02
0.5680419E-01	0.1124140E+00	0.1680696E-01	0.9691268E-01	-0.1273040E+00
-0.1853897E-03	-0.3199223E-03	-0.4594649E-04	-0.2210247E-03	0.2391982E-03

Poly(May)	Poly(June)	Poly(July)	Poly(August)	Poly(September)
-0.2301289E+00	-0.3631833E+00	-0.4364970E+00	-0.3678722E+00	-0.4959048E+00
0.5305769E+00	0.6408592E+00	0.6707261E+00	0.6470548E+00	0.5209985E+00
0.3176902E+00	0.3698111E+00	0.3778238E+00	0.4291340E+00	0.5681123E+00
0.3939341E+00	0.4674823E+00	0.5361906E+00	0.5193311E+00	0.5062233E+00
0.5062279E+00	0.3246658E+00	0.3803952E+00	0.5951137E+00	0.5703800E+00
0.3771116E+00	0.1280630E+00	-0.3222886E-01	-0.3374881E+00	0.8640252E+00

APPENDIX D. INPUT CONSTANTS FOR LOGISTIC NPTI

F(X1) May	F(X1) June	F(X1) July	F(X1) August	F(X1) September
-0.245742152E+1	-0.874825903E+0	-0.125785528E+0	-0.241378075E+0	-0.875096225E+0
-0.175571837E+0	-0.108233194E+0	-0.136059651E+0	-0.112923323E+0	-0.845048228E-1
-0.377914846E-1	-0.605543392E-1	0.554984982E-2	0.585889123E-2	-0.197709309E-1
0.108514491E-2	-0.866053496E-3	-0.126541745E-2	0.414499420E-2	-0.417157588E-3
-0.447443648E-2	-0.221838304E-2	-0.273152413E-2	-0.110424454E-2	-0.134699595E-2
0.167817244E-2	0.207449860E-2	-0.184289137E-2	-0.227515840E-2	-0.223001017E-2
-0.375266585E-4	0.523305038E-4	0.908985678E-4	0.123380071E-3	0.102828663E-3
0.602941212E-4	-0.712906654E-4	-0.477644243E-4	-0.358966390E-4	-0.156920183E-3
0.199788826E-3	0.247086324E-4	-0.116388478E-3	0.228944231E-3	0.211764502E-4
-0.928873549E-4	0.653953357E-4	0.436216815E-6	-0.892181291E-4	-0.221632376E-4

F(X2) May	F(X2) June	F(X2) July	F(X2) August	F(X2) September
-0.112808872E+1	-0.936047050E+0	-0.691599188E+0	-0.459351617E+0	-0.741845629E+0
-0.122880696E+0	-0.162072255E+0	-0.188523902E+0	-0.134189574E+0	-0.736043120E-1
-0.365253507E-1	-0.985218517E-1	-0.102000033E+0	-0.564064289E-1	-0.771094900E-1
0.153282128E-2	-0.477099398E-2	-0.915040098E-3	0.150876188E-2	0.396630824E-2
-0.812714790E-2	-0.242612182E-2	0.385324800E-3	0.813519706E-3	-0.289976281E-2
0.862859923E-3	0.151235824E-3	-0.269224384E-2	-0.194646161E-2	-0.254828753E-2
-0.137977701E-3	0.186775455E-3	0.288908716E-3	0.233551088E-3	0.107212321E-3
0.621995635E-4	-0.188589110E-3	0.211414286E-3	0.303933232E-4	-0.110120140E-3
0.209652551E-3	0.286149798E-4	-0.345330314E-4	-0.126352118E-4	0.170345705E-3
-0.890941803E-4	0.927775530E-4	0.114394286E-3	0.485406835E-4	-0.662082877E-5

F(X3) May	F(X3) June	F(X3) July	F(X3) August	F(X3) September
-0.744794227E+1	-0.344504077E+1	-0.339693534E+1	-0.387478991E+1	-0.198661296E+2
0.207507920E+2	-0.412995409E+1	0.266573401E+1	0.673636499E+1	0.725721396E+2
-0.202790761E+2	0.289025740E+2	0.829061265E+1	-0.353171058E+0	-0.929219261E+2
0.707965717E+1	-0.226658267E+2	-0.703647151E+1	-0.226668992E+1	0.408593202E+2

F(X4) May	F(X4) June	F(X4) July	F(X4) August	F(X4) September
-0.572143733E+0	-0.879854110E-1	-0.740427130E-1	-0.237429897E+0	-0.708225493E+0
-0.265854007E+0	-0.223271178E+0	-0.211274506E+0	-0.197372491E+0	-0.300976815E+0
-0.195574097E-1	-0.335588679E-1	-0.898311775E-2	0.871663496E-2	0.165095989E-1

F(X5) May	F(X5) June	F(X5) July	F(X5) August	F(X5) September
-0.374463884E+0	-0.205195233E+0	-0.345612993E+0	-0.327403324E+1	0.570336424E+1
0.937591149E-2	0.233530223E-1	0.103337183E-1	0.166762482E+0	-0.141837861E+0
-0.202255210E-3	0.219217299E-4	-0.246551690E-3	-0.213724215E-2	0.724549114E-3

Poly(May)	Poly(June)	Poly(July)	Poly(August)	Poly(Semptember)
-0.166123602E+2	-0.702350958E+1	-0.563431384E+1	-0.513263981E+1	-0.350185959E+1
0.380953959E+1	0.461139118E+1	0.424573142E+1	0.383315244E+1	0.460598495E+1
0.284121773E+1	0.220907802E+1	0.187438950E+1	0.184741373E+1	0.153949334E+1
0.386390824E+1	0.247766016E+1	0.314187661E+1	0.250967823E+1	0.485909819E+1
0.413317458E+1	0.403659792E+1	0.274126511E+1	0.304620980E+1	0.232653736E+1
0.274441936E+2	0.169227621E+1	0.195046844E+0	0.156779011E+3	-0.282196658E+1

APPENDIX E. MATHCAD® TEMPLATE FOR FORECAST VERIFICATION

ORIGIN≡1

Creates contingency table data in vector format

(V - Observed   U - Forecast)

$$\text{Contingency}(V,U) := \begin{vmatrix} \text{Counting}_4 \leftarrow 0 \\[6pt] \text{for } j \in 1..\,\text{rows}(V) \\[4pt] \quad \begin{vmatrix} \text{Counting}_1 \leftarrow \text{Counting}_1 + 1 & \text{if } \left(V_j = U_j\right)\cdot\left(V_j = 1\right) \\[4pt] \text{Counting}_2 \leftarrow \text{Counting}_2 + 1 & \text{if } \left(V_j < U_j\right) \\[4pt] \text{Counting}_3 \leftarrow \text{Counting}_3 + 1 & \text{if } \left(V_j > U_j\right) \\[4pt] \text{Counting}_4 \leftarrow \text{Counting}_4 + 1 & \text{if } \left(V_j = U_j\right)\cdot\left(V_j = 0\right) \end{vmatrix} \\[6pt] \text{Counting}_5 \leftarrow \Sigma\text{Counting} \\[4pt] \text{Counting} \end{vmatrix}$$

V is vector of 1's and 0's representing observed occurrenc or nonoccurrence of an event.

U is vector of 1's and 0's representing forecast occurrence or nonoccurrence of an event.

Example Data

Observed data $x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$ Forecast data $y = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

$T := \text{Contingency}(x,y)$

Creates contigency table in Table format

$$T = \begin{bmatrix} 3 \\ 1 \\ 0 \\ 1 \\ 5 \end{bmatrix}$$
Yes Forecast & Observed
Yes Forecast & No Observed
No Forecast & Yes Observed
No Forecast & Observed
Total # of Events

$$n(V) := \begin{vmatrix} n_{1,1} \leftarrow V_1 \\[4pt] n_{1,2} \leftarrow V_2 \\[4pt] n_{2,1} \leftarrow V_3 \\[4pt] n_{2,2} \leftarrow V_4 \\[4pt] n \end{vmatrix}$$

$$T2 := n(T) \qquad T2 = \begin{bmatrix} 3 & 1 \\ 0 & 1 \end{bmatrix} \begin{array}{l} \text{Yes} \\ \text{No} \end{array} \text{Forecast}$$

Observed
Yes No

Convert percentages to Yes(1)/No(0) forecast

$$\text{convert}(V,\text{cutoff}) := \begin{vmatrix} P_1 \leftarrow 0 \\[4pt] \text{for } j \in 1..\,\text{rows}(V) \\[4pt] \quad \begin{vmatrix} P_j \leftarrow 1 & \text{if } V_j \geq \text{cutoff} \\[4pt] P_j \leftarrow 0 & \text{if } V_j < \text{cutoff} \end{vmatrix} \\[6pt] P \end{vmatrix}$$

$$p = \begin{bmatrix} 0.5 \\ 0.3 \\ 0.49 \\ 0.6 \\ 0.9 \end{bmatrix} \qquad \text{convert}(p,.5) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

Hit Rate

$$HR(V) := \frac{V_1 + V_4}{V_5} \cdot 100$$

$$HR(T) = 80$$

Threat Score Yes/No

$$TSY(V) := \frac{V_1}{V_1 + V_2 + V_3} \cdot 100$$

$$TSY(T) = 75$$

$$TSN(V) := \frac{V_4}{V_4 + V_2 + V_3} \cdot 100$$

$$TSN(T) = 50$$

Probability of Detection Yes/No

$$PODY(V) := \frac{V_1}{V_1 + V_3} \cdot 100$$

$$PODY(T) = 100$$

$$PODN(V) := \frac{V_4}{V_4 + V_2} \cdot 100$$

$$PODN(T) = 50$$

False Alarm Rate Yes/No

$$FARY(V) := \frac{V_2}{V_1 + V_2} \cdot 100$$

$$FARY(T) = 25$$

$$FARN(V) := \frac{V_3}{V_4 + V_3} \cdot 100$$

$$FARN(T) = 0$$

Forecast Bias

$$BIAS(V) := \frac{V_1 + V_2}{V_1 + V_3} \cdot 100$$

$$BIAS(T) = 133.333$$

Heidke Skill Score

$$HSS(V) := \frac{2 \cdot \left(V_1 \cdot V_4 - V_2 \cdot V_3\right)}{\left(V_1 + V_3\right) \cdot \left(V_3 + V_4\right) + \left(V_1 + V_2\right) \cdot \left(V_2 + V_4\right)}$$

$$HSS(T) = 0.545$$

Kuipers Skill Score

$$KSS(V) := \frac{\left(V_1 \cdot V_4 - V_2 \cdot V_3\right)}{\left(V_1 + V_3\right) \cdot \left(V_2 + V_4\right)}$$

$$KSS(T) = 0.5$$

Chi squared

Note: V is contingency table data in vector format

$$\chi_{sq}(V) := \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{\left[(n(V))_{i,j} - (E(V))_{i,j}\right]^2}{(E(V))_{i,j}}$$

$\chi_{sq}(T) = 1.875$

Perfect Chi ^{2}

$\chi_{sq}(\text{Perf}) = 5$

Persistence Chi ^{2}

$\chi_{sq}(\text{Pst}) = 0.139$

Brier Score

$$BS(V,P) := \frac{100}{\text{rows}(V)} \cdot \sum_{k=1}^{\text{rows}(V)} \left(\cdot - \cdot \right)^2$$

Note: V is the vector of observed values and P is the vector of forecast probabilities for each observed event

$BS(x,p) = 15.002$

Ratio Skill Score

$$RSS(V,P,\rho) := \frac{BS(V,\rho) - BS(V,P)}{BS(V,\rho)} \cdot 100$$

Note: V is the vector of observed values, P is the vector of forecast probabilities for each observed event, and r is the vector for the persistence forecast

$RSS(x,p,P) = 74.997$

Skill Score (c - accuracy measure)

$$SS(c,P,\text{Perf},\rho) := \frac{c(P) - c(\rho)}{c(\text{Perf}) - c(\rho)} \cdot 100$$

Note: c is the accuracy measure function being measured, P is the contingency table data for the forecast in vector format, Perf is the contingency table data for the perfect forecast in vector format, and r is the contingency table data for the persistence forecast in vector format.

$SS(HR,T,\text{Perf},\text{Pst}) = 66.7$

Perfect Scores	Persistence Scores	Forecast Scores	Skill score of each accuracy measure
HR(Perf) = 100	HR(Pst) = 40	HR(T) = 80	SS(HR,T,Perf,Pst) = 66.7
TSY(Perf) = 100	TSY(Pst) = 25	TSY(T) = 75	SS(TSY,T,Perf,Pst) = 66.7
TSN(Perf) = 100	TSN(Pst) = 25	TSN(T) = 50	SS(TSN,T,Perf,Pst) = 33.3
PODY(Perf) = 100	PODY(Pst) = 33.3	PODY(T) = 100	SS(PODY,T,Perf,Pst) = 100
PODN(Perf) = 100	PODN(Pst) = 50	PODN(T) = 50	SS(PODN,T,Perf,Pst) = 0
FARY(Perf) = 0	FARY(Pst) = 50	FARY(T) = 25	SS(FARY,T,Perf,Pst) = 50
FARN(Perf) = 0	FARN(Pst) = 66.667	FARN(T) = 0	SS(FARN,T,Perf,Pst) = 100
BIAS(Perf) = 100	BIAS(Pst) = 66.667	BIAS(T) = 133.333	
HSS(Perf) = 1	HSS(Pst) = ⁻0.154	HSS(T) = 0.545	SS(HSS,T,Perf,Pst) = 60.6
KSS(Perf) = 1	KSS(Pst) = ⁻0.167	KSS(T) = 0.5	SS(KSS,T,Perf,Pst) = 57.1

79

APPENDIX F. BISERIAL CORRELATION RESULTS

	row#	V	V^2	V^3	$V+V^2$	$V+V^3$	V^2+V^3	$V+V^2+V^3$
1000mb								
Heights	31	-0.0984	-0.1143	-0.1242	-0.1143	-0.1242	-0.1242	-0.1242
Temperature	32	0.0699	0.0645	0.0594	0.0646	0.0594	0.0595	0.0595
Dewpoint	33	0.1786	0.1741	0.1675	0.1743	0.1675	0.1678	0.1678
Relative Humidity	34	0.1510	0.1508	0.1495	0.1510	0.1504	0.1502	0.1506
U-wind component	35	-0.2632	-0.0774	-0.0316	-0.0958	-0.0320	-0.0337	-0.0341
V-wind component	36	-0.1987	0.0053	-0.0048	-0.0061	-0.0049	-0.0047	-0.0049
950mb								
Heights	37	-0.0801	-0.0838	-0.0872	-0.0838	-0.0872	-0.0872	-0.0872
Temperature	38	0.1845	0.1874	0.1856	0.1874	0.1856	0.1857	0.1857
Dewpoint	39	0.1748	0.1760	0.1715	0.1761	0.1715	0.1717	0.1718
Relative Humidity	40	0.0659	0.0594	0.0539	0.0621	0.0586	0.0565	0.0589
U-wind component	41	-0.2936	-0.0233	-0.0983	-0.0438	-0.0989	-0.0986	-0.0992
V-wind component	42	-0.1801	0.0254	-0.0485	0.0153	-0.0488	-0.0488	-0.0491
900mb								
Heights	43	-0.0537	-0.0557	-0.0576	-0.0557	-0.0576	-0.0576	-0.0576
Temperature	44	0.1922	0.1900	0.1864	0.1901	0.1864	0.1865	0.1865
Dewpoint	45	0.2085	0.2074	0.1975	0.2077	0.1976	0.1981	0.1981
Relative Humidity	46	0.1049	0.0927	0.0808	0.0980	0.0909	0.0866	0.0916
U-wind component	47	-0.3026	-0.0159	-0.0617	-0.0356	-0.0622	-0.0625	-0.0630
V-wind component	48	-0.2031	0.0282	-0.0426	0.0175	-0.0429	-0.0427	-0.0430
850mb								
Heights	49	-0.0282	-0.0295	-0.0308	-0.0295	-0.0308	-0.0308	-0.0308
Temperature	50	0.1519	0.1466	0.1398	0.1467	0.1398	0.1401	0.1401
Dewpoint	51	0.2338	0.2314	0.2196	0.2332	0.2198	0.2207	0.2209
Relative Humidity	52	0.1784	0.1665	0.1516	0.1724	0.1648	0.1592	0.1654
U-wind component	53	-0.3066	-0.0325	-0.0561	-0.0532	-0.0567	-0.0576	-0.0581
V-wind component	54	-0.2140	0.0118	-0.0408	-0.0001	-0.0411	-0.0412	-0.0415
800mb								
Heights	55	-0.0075	-0.0085	-0.0095	-0.0085	-0.0095	-0.0095	-0.0095
Temperature	56	0.0885	0.0772	0.0653	0.0777	0.0654	0.0659	0.0660
Dewpoint	57	0.2579	0.1688	0.2088	0.1852	0.2093	0.2121	0.2125
Relative Humidity	58	0.2245	0.2075	0.1867	0.2164	0.2071	0.1977	0.2073
U-wind component	59	-0.2956	-0.0443	-0.0487	-0.0637	-0.0492	-0.0505	-0.0510
V-wind component	60	-0.2119	-0.0067	-0.0388	-0.0189	-0.0391	-0.0396	-0.0399
750mb								
Heights	61	-0.0072	-0.0081	-0.0089	-0.0081	-0.0089	-0.0089	-0.0089
Temperature	62	0.0085	-0.0102	-0.0259	-0.0093	-0.0258	-0.0250	-0.0249
Dewpoint	63	0.2755	-0.0441	0.1620	-0.0133	0.1629	0.1655	0.1664
Relative Humidity	64	0.2743	0.2616	0.2407	0.2691	0.2615	0.2522	0.2616
U-wind component	65	-0.2825	-0.0316	-0.0488	-0.0495	-0.0493	-0.0505	-0.0509
V-wind component	66	-0.1950	-0.0098	-0.0340	-0.0202	-0.0342	-0.0342	-0.0344

700mb	row#	V	V²	V³	V+V²	V+V³	V²+V³	V+V²+V³
Heights	67	0.0074	0.0067	0.0060	0.0067	0.0060	0.0060	0.0060
Temperature	68	-0.0407	-0.0654	-0.0807	-0.0639	-0.0805	-0.0796	-0.0794
Dewpoint	69	0.2892	-0.1849	0.1616	-0.1726	0.1622	0.1600	0.1607
Relative Humidity	70	0.3068	0.2965	0.2765	0.3032	0.2970	0.2879	0.2969
U-wind component	71	-0.2649	-0.0099	-0.0384	-0.0253	-0.0388	-0.0394	-0.0398
V-wind component	72	-0.1820	-0.0028	-0.0376	-0.0125	-0.0378	-0.0378	-0.0380
650mb								
Heights	73	-0.0026	-0.0032	-0.0039	-0.0032	-0.0039	-0.0039	-0.0039
Temperature	74	-0.0636	-0.0987	-0.1026	-0.0955	-0.1021	-0.1024	-0.1020
Dewpoint	75	0.2888	-0.2520	0.2057	-0.2484	0.2060	0.2031	0.2035
Relative Humidity	76	0.3029	0.2821	0.2550	0.2939	0.2844	0.2701	0.2837
U-wind component	77	-0.2467	0.0063	-0.0301	-0.0059	-0.0303	-0.0305	-0.0308
V-wind component	78	-0.1769	-0.0039	-0.0527	-0.0139	-0.0529	-0.0530	-0.0532
600mb								
Heights	79	0.0031	0.0026	0.0020	0.0026	0.0020	0.0020	0.0020
Temperature	80	-0.0588	-0.1011	-0.0011	-0.1194	-0.0036	-0.0160	-0.0182
Dewpoint	81	0.2714	-0.2477	0.2101	-0.2461	0.2103	0.2083	0.2085
Relative Humidity	82	0.2863	0.2687	0.2453	0.2791	0.2715	0.2585	0.2706
U-wind component	83	-0.2300	0.0150	-0.0298	0.0039	-0.0300	-0.0300	-0.0303
V-wind component	84	-0.1590	-0.0029	-0.0592	-0.0116	-0.0593	-0.0595	-0.0597
550mb								
Heights	85	-0.0019	-0.0024	-0.0030	-0.0024	-0.0030	-0.0030	-0.0030
Temperature	86	-0.0246	-0.0161	0.0392	-0.0209	0.0386	0.0416	0.0409
Dewpoint	87	0.2420	-0.2263	0.1881	-0.2255	0.1882	0.1866	0.1867
Relative Humidity	88	0.2508	0.2299	0.2062	0.2419	0.2341	0.2194	0.2327
U-wind component	89	-0.2082	0.0025	-0.0174	-0.0068	-0.0176	-0.0177	-0.0179
V-wind component	90	-0.1395	-0.0058	-0.0572	-0.0134	-0.0573	-0.0571	-0.0572
500mb								
Heights	91	0.0032	0.0027	0.0022	0.0027	0.0022	0.0022	0.0022
Temperature	92	0.0092	-0.0276	0.0424	-0.0286	0.0423	0.0434	0.0433
Dewpoint	93	0.1989	-0.1821	0.1357	-0.1816	0.1358	0.1345	0.1346
Relative Humidity	94	0.2044	0.1847	0.1630	0.1961	0.1891	0.1752	0.1877
U-wind component	95	-0.1707	-0.0161	-0.0086	-0.0232	-0.0087	-0.0091	-0.0092
V-wind component	96	-0.1339	-0.0340	-0.0313	-0.0388	-0.0313	-0.0314	-0.0315

Thunderstorm Indices	row#	V	V^2	V^3	$V+V^2$	$V+V^3$	V^2+V^3	$V+V^2+V^3$
SSI	97	-0.2906	-0.1912	-0.1574	-0.2059	-0.1591	-0.1605	-0.1621
LI	98	-0.2616	0.0816	-0.1692	0.0256	-0.1725	-0.1589	-0.1626
K	99	0.3423	0.3603	0.3580	0.3603	0.3580	0.3581	0.3581
VT	100	0.1441	0.1375	0.1304	0.1377	0.1304	0.1306	0.1306
CT	101	0.2518	0.2652	0.2635	0.2650	0.2635	0.2637	0.2637
TT	102	0.2766	0.2797	0.2789	0.2797	0.2789	0.2790	0.2790
TI	103	0.3521	0.3710	0.3648	0.3711	0.3648	0.3651	0.3651
MDPI	104	0.1713	0.1697	0.1681	0.1698	0.1682	0.1682	0.1682
CCL	105	0.1161	0.1220	0.1200	0.1220	0.1200	0.1200	0.1200
LCL	106	0.1474	0.1478	0.1482	0.1478	0.1482	0.1482	0.1482
LCL-CCL	107	0.1474	0.1478	0.1482	0.1478	0.1482	0.1482	0.1482
RH 1000 to 700mb	108	0.2934	0.2905	0.2740	0.2919	0.2742	0.2752	0.2754
RH 800 to 600mb	109	0.2533	0.2538	0.2450	0.2540	0.2450	0.2455	0.2456
Thickness								
500mb	110	0.0710	0.0707	0.0704	0.0707	0.0704	0.0704	0.0704
700mb	111	0.1381	0.1379	0.1376	0.1379	0.1376	0.1376	0.1376
850mb	112	0.1455	0.1450	0.1446	0.1450	0.1446	0.1446	0.1446
800 to 500mb	113	0.0362	0.0357	0.0353	0.0357	0.0353	0.0353	0.0353
800 to 700mb	114	0.1037	0.1033	0.1030	0.1033	0.1030	0.1030	0.1030
700 to 500mb	115	-0.0061	-0.0066	-0.0071	-0.0066	-0.0071	-0.0071	-0.0071
Vertical Wind Shear								
1000 to 500mb	116	-0.0841	-0.0753	-0.0510	-0.0756	-0.0510	-0.0514	-0.0515
700 to 500mb	117	-0.1076	-0.0928	-0.0605	-0.0935	-0.0606	-0.0613	-0.0614
700 to 1000mb	118	-0.0730	-0.0783	-0.0580	-0.0784	-0.0580	-0.0586	-0.0586
850 to 600mb	119	-0.0880	-0.0853	-0.0735	-0.0856	-0.0735	-0.0739	-0.0740

This program is a function of an array(include the patterns of persistence), the length of pattern to look at, and a day number withholding function. It loops through the decimal equivalent of the binary patterns of persistence and increments a thunderstorm counter or a no thunderstorm counter for the pattern observed for that day. The next day is read and the loop starts again.

The output is an array holding the number times a thunderstorm follows each pattern by month and how many times no thunderstorm follows each pattern by month.

<u>Function to withhold day number of missing years</u>
(uses lightning data)(x is day number)

$$La(x) := ((x < 5048) + (x > 5355)) \cdot ((x < 4285) + (4896 < x))$$

$Count(V, n, ex) :=$

$$day_{2^n \cdot 2, 5} \leftarrow 0$$

$r \leftarrow cols(V)$

$d \leftarrow 1$

$m \leftarrow 1$

for $i \in 1 .. rows(V)$

    for $j \in 1 .. 2^n$

        if $\left(V_{i,r} = j - 1\right) \cdot \left(ex\left(V_{i,2}\right) = 1\right) \cdot \left(V_{i,3} > n\right) \cdot \left(V_{i,3} \leq 153\right) \cdot \left(V_{i,1} = 1\right)$

            $day_{j,m} \leftarrow day_{j,m} + 1$

            day

        if $\left(V_{i,r} = j - 1\right) \cdot \left(ex\left(V_{i,2}\right) = 1\right) \cdot \left(V_{i,3} > n\right) \cdot \left(V_{i,3} \leq 153\right) \cdot \left(V_{i,1} = 0\right)$

            $day_{j + 2^n, m} \leftarrow day_{j + 2^n, m} + 1$

            day

    if $((m = 1) \cdot (d = 30)) + (d = 31)$

        $d \leftarrow 0$

        $m \leftarrow m + 1$

    if $V_{i,3} = 1$

        $m \leftarrow 1$

        $d \leftarrow 0$

    $d \leftarrow d + 1$

day

This program is a function of an array(created by Count program) and the length of pattern to look at. It loops through the decimal equivalent of the binary patterns of persistence and calculates the conditional probability from the relative frequencies. The output is an array holding the conditional probability of a thunderstorm for each pattern.

$$\text{NextDayProbyMon}(V, n) := \begin{vmatrix} \text{answer}_{1,1} \leftarrow 0 \\ \text{for } h \in 1 .. 5 \\ \quad \text{for } i \in 1 .. 2^n \\ \qquad \text{answer}_{i,h} \leftarrow \dfrac{V_{i,h}}{V_{i,h} + V_{i+2^n,h}} \\ \text{answer} \end{vmatrix}$$

Example Result
Probability of a Thunderstorm given a
3 Day Persistence Pattern

	May	Jun	Jul	Aug	Sep
000	0.132	0.235	0.281	0.272	0.193
001	0.260	0.264	0.282	0.333	0.228
010	0.220	0.325	0.386	0.340	0.195
011	0.232	0.398	0.333	0.370	0.258
100	0.414	0.552	0.695	0.590	0.542
101	0.484	0.606	0.587	0.567	0.605
110	0.393	0.643	0.628	0.508	0.473
111	0.462	0.654	0.727	0.690	0.520

Example Result
3 Day Pattern Frequencies
(number of pattern occurrences)

Pattern	May	Jun	Jul	Aug	Sep	
000	87	100	95	85	102	Patterns Preceding Thunderstorms
001	27	33	37	49	29	
010	18	27	27	32	15	
011	13	45	42	51	24	
100	48	74	89	79	71	
101	15	43	37	51	23	
110	24	74	81	66	44	
111	18	134	208	156	52	
000	572	325	243	227	426	Patterns Preceding no Thunderstorms
001	77	92	94	98	98	
010	64	56	43	62	62	
011	43	68	84	87	69	
100	68	60	39	55	60	
101	16	28	26	39	15	
110	37	41	48	64	49	
111	21	71	78	70	48	

APPENDIX H. MATHCAD® TEMPLATE FOR
CREATION OF WIND SECTORS

This program is a function of data set array and level of wind to be correlated.
The u and v components of the wind are converted to direction for every level.
A variable is set to 1 or 0 depending on whether the wind falls within the
specified sector. Every level is determined. The output is an array of 1's and 0's
with each column representing a different level. These columns can then be
correlated to thunderstorm occurrence.

$\text{WndSector(V, level)} :=$

$$\left|\begin{array}{l} \text{ans}_{\text{rows}(V),12} \leftarrow 0 \\[4pt] \text{for } i \in 1..11 \\[4pt] \quad \text{for } j \in 1..\text{rows}(V) \\[4pt] \qquad \left|\begin{array}{l} \text{dir} \leftarrow \left|\begin{array}{l} 0 \quad \text{if } V_{j,(\text{level})} = 0 \\[6pt] \left[\left[\dfrac{\pi}{2} - \text{atan}\left[\dfrac{V_{j,(\text{level}+1)}}{V_{j,(\text{level})}}\right]\right] + \pi\cdot\left[V_{j,(\text{level})} < 0\right]\right]\cdot\dfrac{180}{\pi} \quad \text{if } \left[V_{j,(\text{level})} \neq 0\right] \end{array}\right. \\[10pt] m \leftarrow 0 \quad \text{if } (((\text{dir} \geq 0)\cdot(\text{dir}<190)) + (\text{dir} \geq 300)) \\[4pt] m \leftarrow 1 \quad \text{if } (\text{dir} \geq 190)\cdot(\text{dir}<300) \\[4pt] \text{ans}_{j,i+1} \leftarrow m \end{array}\right. \\[6pt] \text{ans} \end{array}\right.$$

APPENDIX I. MATHCAD® TEMPLATE FOR
BISERIAL CORRELATION

This program is a function of data set array and column of variable to be correlated. It loops through the array, calculates the average values of the variable with and without thunderstorm occurrence, and counts the number of occurrences and nonoccurrences. It uses these numbers to calculate the point biserial correlation coeffecient.

$\text{biserial}(V, c) :=$

$sX_3 \leftarrow 0$

$s_3 \leftarrow 0$

$S \leftarrow 0$

for $i \in 1 .. \text{rows}(V)$

  $sX_1 \leftarrow sX_1 + V_{i,c}$

  $s_1 \leftarrow i$

  if $V_{i,1} = 1$

    $sX_2 \leftarrow sX_2 + V_{i,c}$ Calculate number of

    $s_2 \leftarrow s_2 + 1$ thunderstorm occurrences

  if $V_{i,1} = 0$

    $sX_3 \leftarrow sX_3 + V_{i,c}$ Calculate number of

    $s_3 \leftarrow s_3 + 1$ thunderstorm non occurrences

for $j \in 1 .. 3$

  $X_j \leftarrow \dfrac{sX_j}{s_j}$ Calculate averages

for $i \in 1 .. \text{rows}(V)$

  $S \leftarrow S + \left(V_{i,c} - X_1\right)^2$

$\text{ans} \leftarrow \sqrt{\dfrac{s_2 \cdot s_3}{s_1}} \cdot \left(\dfrac{X_2 - X_3}{\sqrt{S}}\right)$ Calculate Correlation

ans

APPENDIX J. ALGORITHM AND INPUT CONSTANTS FOR STRATIFIED LOGISTIC THUNDERSTORM INDEX

ORIGIN ≡ 1

This algorithm takes an array of variables and the coefficients determined by regression and outputs and array of probabilities representing the probability of thunderstorm occurrence.

Final(V, FC, NE, SW, coeff, g1, g2, g3, g4, g5) :=

$w \leftarrow (20\ 24\ 103\ 99\ 109)$

$lev \leftarrow (53\ 71\ 83)$

for $j \in 1 .. rows(V)$

$m \leftarrow ceil \left[\dfrac{\left(V_{j,2} - 153 \cdot floor\left(\dfrac{V_{j,2}-1}{153}\right) \right) + ceil\left(\dfrac{V_{j,2} - 153 \cdot floor\left(\dfrac{V_{j,2}-1}{153}\right) + 1}{62} \right) - 1}{31} \right]$ determines what month it is

$lev2 \leftarrow 47$ if $(m=1) + (m=5)$

$lev2 \leftarrow 65$ if $(m=2)$

$lev2 \leftarrow 59$ if $(m=3) + (m=4)$ defines the wind level to use

$dir \leftarrow \begin{vmatrix} 0 & \text{if } V_{j,lev2}=0 \\ \left[\left[\dfrac{\pi}{2} - atan\left(\dfrac{V_{j,lev2+1}}{V_{j,lev2}}\right) + \pi \cdot (V_{j,lev2}<0)\right] \cdot \dfrac{180}{\pi}\right] & \text{if } V_{j,lev2} \ne 0 \end{vmatrix}$

determines wind direction from u and v wind components

$mdir \leftarrow 1$ if $(((dir<160) + (dir \ge 300)) \cdot m=1)$

$mdir \leftarrow 1$ if $(((dir<190) + (dir \ge 310)) \cdot m=2)$

$mdir \leftarrow 1$ if $(((dir<180) + (dir \ge 300)) \cdot m=3)$

$mdir \leftarrow 1$ if $(((dir<190) + (dir \ge 300)) \cdot m=4)$ defines the wind sector to use

$mdir \leftarrow 1$ if $(((dir<160) + (dir \ge 310)) \cdot m=5)$

$mdir \leftarrow 0$ if $((dir \ge 160) \cdot (dir<300) \cdot m=1)$

$mdir \leftarrow 0$ if $((dir \ge 190) \cdot (dir<310) \cdot m=2)$

$mdir \leftarrow 0$ if $((dir \ge 180) \cdot (dir<300) \cdot m=3)$

$mdir \leftarrow 0$ if $((dir \ge 190) \cdot (dir<300) \cdot m=4)$

$mdir \leftarrow 0$ if $((dir \ge 160) \cdot (dir<310) \cdot m=5)$

for $i \in 1 .. 5$

   if $mdir=1$

     for $i \in 1 .. 5$

      $temp \leftarrow g4\left[V_{j,(w_{1,i})}, coeff, m, i\right]$

      $ans_i \leftarrow \dfrac{e^{temp}}{1 + e^{temp}}$

     1

calculates level 1 regression equation results of the 1st five variables (all but winds) for the NE sector

88

$$1$$

$$temp \leftarrow g1\left(V_{j,lev_{1,1}}, V_{j,lev_{1,1}+1}, NE, m\right)$$

$$ans_6 \leftarrow \frac{e^{temp}}{1 + e^{temp}}$$

    calculates level 1 regression equation results for NE 850-mb winds

$$temp \leftarrow g2\left(V_{j,lev_{1,2}}, V_{j,lev_{1,2}+1}, NE, m\right)$$

$$ans_7 \leftarrow \frac{e^{temp}}{1 + e^{temp}}$$

    calculates level 1 regression equation results for NE 700-mb winds

$$temp \leftarrow g3\left(V_{j,lev_{1,3}}, V_{j,lev_{1,3}+1}, NE, m\right)$$

$$ans_8 \leftarrow \frac{e^{temp}}{1 + e^{temp}}$$

    calculates level 1 regression equation results for NE 600-mb winds

if mdir=0

    for $i \in 1..5$

$$temp \leftarrow g4\left[V_{j,\left(w_{1,i}\right)}, coeff, m, i+5\right]$$

$$ans_i \leftarrow \frac{e^{temp}}{1 + e^{temp}}$$

    calculates level 1 regression equation results of the 1st five variables (all but winds) for the SW sector

$$temp \leftarrow g1\left(V_{j,lev_{1,1}}, V_{j,lev_{1,1}+1}, SW, m\right)$$

$$ans_6 \leftarrow \frac{e^{temp}}{1 + e^{temp}}$$

    calculates level 1 regression equation results for SW 850-mb winds

$$temp \leftarrow g2\left(V_{j,lev_{1,2}}, V_{j,lev_{1,2}+1}, SW, m\right)$$

$$ans_7 \leftarrow \frac{e^{temp}}{1 + e^{temp}}$$

    calculates level 1 regression equation results for SW 700-mb winds

$$temp \leftarrow g3\left(V_{j,lev_{1,3}}, V_{j,lev_{1,3}+1}, SW, m\right)$$

$$ans_8 \leftarrow \frac{e^{temp}}{1 + e^{temp}}$$

    calculates level 1 regression equation results for SW 600-mb winds

$$result_{j,1} \leftarrow V_{j,1}$$

$$result_{j,2} \leftarrow V_{j,5}$$

$$temp \leftarrow g5\left(ans_1, ans_2, ans_3, ans_4, ans_5, ans_6, ans_7, ans_8, FC, m\right)$$

$$result_{j,3} \leftarrow \frac{e^{temp}}{1 + e^{temp}}$$

    calculates level 2 regression equation results

result

Coefficients for Level 1 regression of 3rd order polynomials (g4)
Each consecutive group of 4 numbers in each column represent the coefficients $g4(a, cf, m, c) := cf_{4,m-3,c} + cf_{4,m-2,c} \cdot a + cf_{4,m-1,c} \cdot a^2 + cf_{4,m,c} \cdot a^3$
for the 3rd order polynomial for that month, variable, and wind sector

Month	NE 6-day persistence	NE 15-day Average	NE Thompson Index	NE K Index	NE 800-mb to 600-mb RH	SW 6-day persistence	SW 15-day Average	SW Thompson Index	SW K Index	SW 800-mb to 600-mb RH
May	-3.30144566	-2.55680267	-20.46528675	-20.6273829	-20.17724673	-0.97664131	-2.68710658	-2.48838997	-4.300094	-3.5836949
	5.69817644	1.69753121	1.47045673	1.69604948	76.23174796	1.3079444	18.24352295	0.13422016	0.41745442	5.69530198
	-2.35660515	0.56821472	-0.03794602	-0.05197621	-105.66732081	1.87801913	-33.36889787	-0.00420034	-0.01735035	2.96677664
	0	0	0.00034455	0.00056917	50.29925028	0	0	0.0000772	0.00027276	-4.93939217
June	-2.02929586	-12.65718735	-48.50124298	-6.37005212	-23.20167922	-2.51069087	4.34390568	-6.24675937	-3.16222879	-1.95077061
	1.36026523	52.67965086	4.23874676	0.10451191	77.96010417	9.25025783	-22.67662385	0.46956815	0.08595428	-6.29725242
	0.53793253	-60.29846311	-0.12648206	0.00585199	-90.03248048	-6.44344271	29.18267139	-0.01339419	0.00185539	29.99987208
	0	0	0.00126296	-0.00011524	34.69865781	0	0	0.00015523	-0.00003274	-22.83263351
July	-5.1976375	526.22533253	-5.13697937	-4.19204576	-0.47879914	-1.93665136	-257.63174604	-5.13697937	-2.43773261	-4.62098534
	12.32136761	-2209.53820727	0.06332135	0.0414967	-19.47258505	6.62512697	1071.117644	0.06332135	0.28205765	17.2774412
	-8.50306595	2309.7235745	0.00289415	0.00255898	45.7045223	-3.08122156	-1110.10369627	0.00289415	-0.01408498	-23.40801168
	0	0	-0.00004477	-0.00003401	-28.00724084	0	0	-0.00004477	0.0002616	13.69392734
August	-2.02554919	-12.96534785	-6.08011816	-4.77520144	-10.6659047	-3.18314622	-13.39888816	-2.75738598	-2.24605053	-2.63163238
	0.29383432	58.68409416	0.37511506	0.29598497	38.50157369	8.97576436	54.42369237	0.15820712	0.13990409	6.12258538
	4.26444137	-70.74205621	-0.01103717	-0.01079857	-53.0235611	-3.275096	-52.93277855	-0.00287723	-0.00331286	-1.67099505
	0	0	0.00013083	0.00016801	25.73369654	0	0	0.00002726	0.00004747	-1.24178441
September	-2.73959472	-0.61196687	-4.87917846	-4.37771229	-21.36135806	-1.21035014	-1.1973601	-7.48253642	-5.35025	-29.08035405
	3.40498944	-12.82208577	-0.00753553	-0.02274319	67.14226764	4.21411135	1.30565099	0.69743333	0.49346607	131.68669625
	1.14685173	31.81529017	0.005107	0.00502243	-77.3828412	-0.78569899	11.83770409	-0.02273063	-0.01616486	-193.10685756
	0	0	-0.00005239	-0.00003852	31.88552801	0	0	0.00026191	0.00019569	92.77673167

1st level regression functions of South West winds

850-mb wind function

$$g1(a,b,cf,m) := cf_{1,m} + cf_{2,m}\cdot a + cf_{3,m}\cdot b + cf_{4,m}\cdot a\cdot b + cf_{5,m}\cdot a^2 + cf_{6,m}\cdot b^2 + cf_{7,m}\cdot a^3 + cf_{8,m}\cdot a^2\cdot b + cf_{9,m}\cdot a\cdot b^2 + cf_{10,m}\cdot b^3$$

700-mb wind function

$$g2(a,b,cf,m) := cf_{11,m} + cf_{12,m}\cdot a + cf_{13,m}\cdot b + cf_{14,m}\cdot a\cdot b + cf_{15,m}\cdot a^2 + cf_{16,m}\cdot b^2 + cf_{17,m}\cdot a^3 + cf_{18,m}\cdot a^2\cdot b + cf_{19,m}\cdot a\cdot b^2 + cf_{20,m}\cdot b^3$$

600-mb wind function

$$g3(a,b,cf,m) := cf_{21,m} + cf_{22,m}\cdot a + cf_{23,m}\cdot b + cf_{24,m}\cdot a\cdot b + cf_{25,m}\cdot a^2 + cf_{26,m}\cdot b^2 + cf_{27,m}\cdot a^3 + cf_{28,m}\cdot a^2\cdot b + cf_{29,m}\cdot a\cdot b^2 + cf_{30,m}\cdot b^3$$

	May	June	July	August	September	
SW =	-0.75919036	-0.84779476	-0.40809706	-0.5277348	-0.52109344	1st ten coefficients of each column are for 850-mb winds
	-0.13746693	-0.15052973	-0.22200973	-0.19402977	-0.11915831	
	-0.03594596	-0.08425752	-0.09686717	-0.08303573	-0.11216879	
	-0.00624943	-0.0032473	-0.01330786	-0.00251802	-0.01291393	
	-0.01048502	-0.00365127	-0.00615273	-0.00350807	0.00186946	
	0.00106757	0.00177423	0.00040953	-0.00451718	-0.00074532	
	-0.00017196	0.00008448	0.00008634	0.00021188	0.00054875	
	-0.00014651	-0.00000376	-0.00019842	-0.00031611	-0.00052938	
	-0.0000142	-0.00004783	-0.00034994	0.00010146	-0.00018399	
	-0.00003904	0.00013329	0.00002408	-0.00003721	0.00011901	
	-1.12567318	-0.66017489	0.22076714	-0.11558421	-0.17808479	
	-0.15739506	-0.16380442	-0.1774006	-0.13462212	-0.08943668	
	0.00432637	-0.04127778	-0.03227301	-0.03459424	-0.05784074	
	0.00372987	0.0017943	-0.00566766	0.00313752	-0.00357767	2nd ten coefficients of each column are for 700-mb winds
	-0.00863015	-0.00708765	-0.00784701	-0.00243253	-0.00422197	
	0.00522078	0.00281135	-0.00035433	-0.00301364	0.00279638	
	-0.00012226	-0.00001356	-0.00002144	0.00011727	-0.00002639	
	0.00005712	0.00013638	-0.00011283	0.00000259	0.00011895	
	0.00027315	-0.00009932	-0.00011724	0.00009431	-0.00015752	
	0.00009121	0.00010967	0.00013858	-0.00001557	0.00018713	
	-1.70812358	-0.49854714	0.4455526	-0.06419189	-0.00433269	
	-0.09810054	-0.14480572	-0.16233696	-0.15369527	-0.04228522	
	-0.00926169	-0.07276583	0.04851311	0.00083738	-0.03132065	
	0.00820946	-0.00405125	-0.00002992	0.00127029	-0.00542831	
	-0.00129416	-0.00423608	-0.00845094	-0.00531093	0.00354932	last ten coefficients of each column are for 600-mb winds
	0.00753021	-0.00050512	-0.0000204	-0.00126907	-0.00196243	
	0.00000875	0.00003869	-0.00006097	-0.0000378	0.00025694	
	0.00023735	-0.00004204	-0.0000786	0.00003954	-0.00033321	
	0.00031963	-0.0001298	0.00004709	0.000082	0.00009964	
	0.00013389	0.00004934	-0.00008114	-0.00002679	-0.00005453	

91

850-mb wind function   1st level regression functions of North East winds

$$g1(a,b,cf,m) := cf_{1,m} + cf_{2,m} \cdot a + cf_{3,m} \cdot b + cf_{4,m} \cdot a \cdot b + cf_{5,m} \cdot a^2 + cf_{6,m} \cdot b^2 + cf_{7,m} \cdot a^3 + cf_{8,m} \cdot a^2 \cdot b + cf_{9,m} \cdot a \cdot b^2 + cf_{10,m} \cdot b^3$$

700-mb wind function

$$g2(a,b,cf,m) := cf_{11,m} + cf_{12,m} \cdot a + cf_{13,m} \cdot b + cf_{14,m} \cdot a \cdot b + cf_{15,m} \cdot a^2 + cf_{16,m} \cdot b^2 + cf_{17,m} \cdot a^3 + cf_{18,m} \cdot a^2 \cdot b + cf_{19,m} \cdot a \cdot b^2 + cf_{20,m} \cdot b^3$$

600-mb wind function

$$g3(a,b,cf,m) := cf_{21,m} + cf_{22,m} \cdot a + cf_{23,m} \cdot b + cf_{24,m} \cdot a \cdot b + cf_{25,m} \cdot a^2 + cf_{26,m} \cdot b^2 + cf_{27,m} \cdot a^3 + cf_{28,m} \cdot a^2 \cdot b + cf_{29,m} \cdot a \cdot b^2 + cf_{30,m} \cdot b^3$$

	May	June	July	August	September	
	−1.39735457	−0.81192688	−0.45935162	−0.5277348	−1.42248341	
	−0.16337932	−0.17903852	−0.13418957	−0.19402977	0.01011239	
	−0.01924393	−0.11318309	−0.05640643	−0.08303573	−0.0153742	
	0.00157409	−0.01017808	0.00150876	−0.00251802	0.0044145	
	−0.00787899	−0.01144756	0.00081352	−0.00350807	−0.00223024	1st ten coefficients of each
NE =	−0.00045066	0.00027117	−0.00194646	−0.00451718	−0.00468089	column are for 850-mb
	0.00037086	0.0006944	0.00023355	0.00021188	0.00005116	winds
	−0.00017156	0.00003373	0.00003039	−0.00031611	−0.00024535	
	0.00036136	0.00000176	−0.00001264	0.00010146	0.00018699	
	−0.00006346	0.0002151	0.00000485	−0.00003721	−0.0000611	
	−1.93562709	−1.50057274	−1.98545234	−0.75533321	−1.35217142	
	−0.17937277	−0.05107418	0.02737546	−0.08707048	−0.02634108	
	−0.05167431	−0.00833508	−0.14228185	−0.01727309	−0.0304689	
	−0.00275655	0.00303799	0.00598971	−0.00344996	−0.00199523	
	−0.00901157	−0.00337095	0.00881725	−0.00097115	−0.00331195	
	0.00005605	0.00509632	−0.00652579	−0.00296889	−0.00128174	
	−0.00010536	0.00009911	−0.00045195	0.00018757	0.00008544	2nd ten coefficients of each
	0.00028875	−0.00028071	0.00065115	0.00036129	0.00006235	column are for 700-mb
	0.00078496	−0.00023966	−0.00067685	0.00015726	0.00006666	winds
	−0.00016811	−0.00013838	0.00012697	−0.00011107	0.00001151	
	−2.65493895	−1.32577775	−1.72384294	−0.94921838	−1.44168343	
	−0.15548608	−0.07542698	0.04864186	−0.07508946	−0.01195839	
	−0.05776902	−0.03263039	−0.08591294	−0.01646786	−0.00765895	
	−0.0023119	0.00317909	−0.00628692	0.0011504	−0.00448007	
	−0.00199548	−0.00503603	0.00104194	−0.00176458	0.00008548	
	0.00099906	−0.00057824	−0.00472537	0.00113819	−0.00451049	
	0.00004846	−0.00004537	−0.00057493	0.00020504	0.00001635	last ten coefficients of each
	0.00000459	0.00006697	0.00000348	−0.00002301	0.00001679	column are for 600-mb
	0.00038321	0.00090595	−0.00115929	0.00000735	−0.00001327	winds
	−0.00002425	−0.0004214	0.00003981	−0.00004252	−0.00016118	

Coefficients for 2nd level regression (function g5)

	May	June	July	August	September
FC =	−3.52422053	−3.19376065	−4.34717717	−2.88225896	−2.88225896
	2.02870833	2.71884464	3.66172082	3.66588934	3.66588934
	−2.36366193	−7.01820778	0.23900732	−0.73233145	−0.73233145
	8.40441166	8.15811753	9.22149719	2.96923916	2.96923916
	−3.30920217	−4.42014698	−3.42312361	1.56811367	1.56811367
	0.02861214	2.17740649	1.44730897	−2.73732854	−2.73732854
	2.52491409	3.50578959	2.62274102	0.48034821	0.48034821
	3.40919978	−0.3646134	−1.20114017	0.52840618	0.52840618
	−2.93856522	1.66948952	2.00008196	0.60340528	0.60340528

2nd level regression function

$$g5(a1,a2,a3,a4,a5,a6,a7,a8,cf,m) := cf_{1,m} + cf_{2,m} \cdot a1 + cf_{3,m} \cdot a2 + cf_{4,m} \cdot a3 + cf_{5,m} \cdot a4 + cf_{6,m} \cdot a5 + cf_{7,m} \cdot a6 + cf_{8,m} \cdot a7 + cf_{9,m} \cdot a8$$

93

APPENDIX K. CLIMATOLOGICAL FREQUENCIES

Climatological Frequencies using 15-day
weighted average for each Day number
1 May is day 1

Day	Value	Day	Value	Day	Value	Day	Value	Day	Value
1 May	0.146	36	0.336	71	0.493	106	0.477	141	0.27
	0.122		0.334	72	0.495		0.486	142	0.272
	0.122		0.33	73	0.493		0.478	143	0.269
	0.073		0.335	74	0.481		0.475		0.255
	0.09		0.345	75	0.47		0.467		0.248
	0.121		0.349	76	0.466		0.469		0.242
	0.132		0.356	77	0.466		0.466		0.233
	0.161		0.352	78	0.468		0.474		0.213
	0.154		0.351	79	0.468		0.461	149	0.259
	0.164		0.359	80	0.464		0.439	150	0.277
	0.174		0.371	81	0.462		0.416		0.317
	0.202		0.381	82	0.458		0.398		0.195
	0.206		0.392	83	0.46		0.382		0.195
	0.227		0.402	84	0.465		0.372		
	0.219		0.412	85	0.464		0.352		
	0.208		0.428	86	0.47		0.336		
Climo =	0.201	Climo =	0.45	Climo =	0.468	Climo =	0.317	Climo =	0.242
	0.202		0.458	88	0.468		0.312		
	0.206		0.468	89	0.466		0.316		
	0.213		0.464	90	0.467		0.327		
	0.224		0.47	91	0.465		0.328		
	0.231		0.47		0.468		0.333		
	0.234		0.476		0.476		0.328		
	0.239		0.479		0.473		0.326		
	0.24		0.48	95	0.472		0.331		
	0.254		0.47	96	0.468		0.335		
	0.266		0.469	97	0.466	132	0.332		
	0.279		0.47	98	0.465	133	0.321		
	0.29		0.477	99	0.474	134	0.306		
	0.293		0.488	100	0.481	135	0.294		
	0.291		0.495	101	0.482	136	0.288		
	0.295		0.497	102	0.479	137	0.281		
	0.303		0.496	103	0.476	138	0.277		
	0.317		0.489	104	0.467	139	0.263		
	0.333		0.495	105	0.471	140	0.276		

APPENDIX L. 6-DAY CONDITIONAL PROBABILITIES

Columns
1 May
2 Jun
3 Jul
4 Aug
5 Sep

000000
000001
000010
000011
.
.
.

a6 =

1 May	2 Jun	3 Jul	4 Aug	5 Sep
0.145	0.19	0.268	0.33	0.191
0.113	0.396	0.277	0.149	0.179
0	0.265	0.292	0.31	0.195
0.12	0.243	0.304	0.375	0.15
0.171	0.29	0.368	0.185	0.188
0.5	0.286	0.083	0.25	0.167
0.2	0.19	0.286	0.2	0.286
0.091	0.167	0.308	0.25	0.24
0.324	0.273	0.222	0.238	0.222
0.167	0.444	0.25	0.231	0.2
0.4	0.125	0.429	0.2	0.25
0	0.25	0.333	0.545	0.2
0.353	0.211	0.273	0.318	0.24
0.2	0	0.308	0.5	0.125
0.333	0.375	0.467	0.143	0.267
0.167	0.304	0.194	0.414	0.263
0.263	0.316	0.353	0.467	0.282
0.167	0.222	0.4	0.25	0
0.1	0.429	0.75	0	0
0.2	0.333	0.222	0.308	0.5
0.429	0.286	0.571	0	0.2
0	0	0.333	0.2	0
0.25	0.5	0.333	0.3	0.167
0	0.444	0.357	0.667	0
0.3	0.316	0.227	0.36	0.2
0.429	0.545	0.2	0.5	0.444
0	0.6	0.8	0.182	0.2
0.5	0.125	0.143	0.278	0.286
0	0.231	0.429	0.308	0.2
0.25	0.333	0.429	0.286	0.4
0.333	0.412	0.231	0.4	0.333
0	0.516	0.385	0.459	0.2

a6 =

1 May	2 Jun	3 Jul	4 Aug	5 Sep
0.422	0.571	0.674	0.537	0.508
0	0.412	0.864	0.625	0.5
0.444	0.615	0.571	0.7	0.625
0.75	0.375	0.684	0.615	0.462
0.167	0.615	0.571	0.571	0.545
1	0.333	0.857	0.625	0.5
0.286	0.25	0.8	0.563	0.75
1	0.769	0.462	0.6	0.75
0.545	0.571	0.5	0.611	0.545
0	0.6	0.6	0	0.5
1	1	0.333	1	1
0.5	0.5	0.5	0.615	1
0.222	0.75	0.6	0.5	0.75
0	0.5	0.833	0.714	0
1	0.8	0.538	0.5	0.8
1	0.522	0.688	0.6	0.571
0.407	0.667	0.625	0.447	0.436
0.571	0.545	0.765	0.733	0.5
0	0.667	0.7	0.778	0.375
0.2	0.5	0.786	0.5	0.583
0.667	0.6	0.75	0.455	0.429
0.5	0.75	0.4	0.444	0.5
0	0.727	0.625	0.333	0.571
1	0.688	0.368	0.5	0.5
0.467	0.719	0.636	0.714	0.478
1	0.833	0.722	0.688	0.6
0.5	0.444	0.875	0.667	0.5
0.5	0.684	0.636	0.714	0.571
0.286	0.636	0.78	0.655	0.6
0.5	0.75	0.692	0.722	0.333
0.667	0.515	0.707	0.758	0.429
0	0.667	0.755	0.658	0.563

BIBLIOGRAPHY

Bauman, William H. III, Michael L. Kaplan, and Steven Businger, 1997: Nowcasting convective activity for space shuttle landings during easterly flow regimes. *Weather and Forecasting*, **12**, 78-107.

Bechtold, Peter, Jean-Pierre Pinty, and Patrick Mascart, 1991: A numerical Investigation of the influence of large-scale winds on sea-breeze- and inland-breeze-type circulations. *Journal of Applied Meteorology*, **30**, 1268-1279.

Byers, Horace R., and Harriet R. Rodebush, 1948: Causes of thunderstorms of the Florida peninsula. *Journal of Meteorology*, **5**, 275-280.

Cooper, H. J., M. Garstang, and J. Simpson,1982: The diurnal interaction between convection and peninsular-scale forcing over south Florida. *Monthly Weather Review*, **110**, 486-503.

Devore, Jay L., 1995: Probability and Statistics for Engineering and the Sciences. 4th ed., Pacific Grove: Duxbury Press, 743 pp.

Dillon, William R., and Matthew Goldstein, 1984: Multivariate Analysis, Methods and Applications. New York: John Wiley & Sons, 587 pp.

Duffield, George F., and G. D. Nastrom, 1983: AWS/TR-83/001, *Equations and Algorithms for Meteorological Applictions in Air Weather Service*. Air Weather Service, 58 pp. Avialable from Defense Technical Information Agency.

Estoque, M.A., 1962: The sea breeze as a function of the prevailing synoptic situation. *Journal of the Atmospheric Sciences*, **19**, 244-250.

Everitt, B. S., 1992: The Analysis of Contingency Tables. 2nd ed., London: Chapman & Hall, 164 pp.

Falls, Lee W., William O. Williford, and Michael C. Carter, 1971: Probability distributions for thunderstorm activity at Cape Kennedy, Florida. *Journal of Applied Meteorology*, **10**, 97-104.

Frank, Neil L., Paul L. Moore, and George E. Fisher, 1967: Summer shower distribution over the Florida peninsula as deduced from digitized radar data. *Journal of Applied Meteorology*, **6**, 309-316.

Gentry, Robert C., and Paul L. Moore, 1954: Relation of local and general wind interactions near the sea coast to time and location of air-mass showers. *Journal of Meteorology*, **11**, 507-511.

Gibbons, Jean D., 1976: <u>Nonparametric Methods for Quantitative Analysis</u>. New York: Holt, Rinehart, & Winston, 463 pp.

Howell, Cindy L., 1998: Nowcasting thunderstorms at Cape Canaveral, Florida, using an improved Neumann-Pfeffer Thunderstorm Index. Unpublished report. The Air Force Institute of Technology, 93 pp.

López, Raúl E., Patrick T. Gannon Sr., David O. Blanchard, and Christopher C. Balch, 1984: Synoptic and regional circulation parameters associated with the degree of convective shower activity in south Florida. *Monthly Weather Review*, **112**, 686-703.

Neter, John., William Wasserman, and Michael H. Kutner, 1983: <u>Applied Linear Regression Models</u>. Illinois: Richard D. Irwin, Inc., 547 pp.

Neumann, Charles J., 1968: Frequency and duration of thunderstorms at Cape Kennedy, part I. Weather Bureau Technical Memorandum SOS-2.

—, 1970: Frequency and duration of thunderstorms at Cape Kennedy, part II. Weather Bureau Technical Memorandum SOS-6.

—, 1971: Thunderstorm forecasting at Cape Kennedy, Florida, utilizing multiple regression techniques. NOAA Technical Memorandum NWS SOS-8.

Nicholls, Melville E., Roger A. Pielke, and William R. Cotton, 1991: A two-dimensional numerical investigation of the interaction between sea breezes and deep convection over the Florida peninsula. *Monthly Weather Review*, **119**, 298-323.

Reed, Jack W., 1979: Cape Canaveral sea breezes. *Journal of Applied Meteorology*, **18**, 231-235.

Roeder, William P., Chief of Operations Support Flight and Science And Technical Training Officer for 45th Weather Squadron, Patrick Air Force Base, Fl. Briefing, February 1998. Personal Interview.

Wilks, D. S., 1995: <u>Statistical Methods in the Atmospheric Sciences</u>. San Diego: Academic Press, 467 pp.

Xian, Zejin, and Roger A. Pielke, 1991: Theeffects of width of landmasses on the development of sea breezes. *Journal of Applied Meteorology*, **30**, 1280-1304.

VITA

James A. Everitt was born in January 1970 in Portage, Michigan, but he was raised in the small town of Crossville, Tennessee. He entered the United States Air Force Academy in June 1988. He graduated and was commissioned in May 1992. His first assignment was to the Basic Meteorology Program at Texas A&M University. After graduating in 1993, his second assignment was to Wright-Patterson AFB as the Wing Weather Officer. In 1996, he was assigned as commander to Detachment 2, 607th Weather Squadron, Camp Humphreys Korea. In August 1997, he was assigned to the Air Force Institute of Technology. His next assignment will be as an instructor at the United States Air Force Academy.

Permanent Address:

206 Ona Rd

Crossville, TN 38555

REPORT DOCUMENTATION PAGE

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE March 1999	3. REPORT TYPE AND DATES COVERED Final

4. TITLE AND SUBTITLE
An Improved Thunderstorm Forecast Index For Cape Canaveral, Florida

5. FUNDING NUMBERS

6. AUTHOR(S)
James A. Everitt, Capt USAF

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
AFIT/ENP
2950 P. Street
Wright-Patterson AFB, OH 45433
Attn: LtCol Cecilia Miner   COMM: (937) 255-3636   DSN: 785-3636
cminer@afit.af.mil

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT/ENP/99M-06

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
45 WS/SYR
1201 Minuteman Street
Patrick AFB, FL 32925-3238
Attn: Mr. William Roeder COMM: (407) 853-8410   DSN: 467-8410
william-roeder@pafb.af.mil

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE
A

13. ABSTRACT *(Maximum 200 words)*
This thesis creates a new algorithm to replace the Neumann-Pfeffer Thunderstorm Index (NPTI). The 45th Weather Squadron at Patrick AFB, Florida, uses the NPTI as an objective means of determining the probability of thunderstorm occurrence at Cape Canaveral. The probability is used for mission planning and resource protection, and increasing the accuracy of NPTI can potentially save billions of dollars for the United States space program. Stratified logistical regressions are performed and probability equations are derived for May through September using upper air data and surface observations for Cape Canaveral. A logistical regression of NPTI was also performed. Variables include combinations of the climatological frequency of thunderstorms, the conditional probability of thunderstorms, the u and v components of the 850-mb, 700-mb, and 600-mb winds, the 800-mb to 600-mb mean relative humidity, the K index, and the Thompson index. The two resulting algorithms are compared to NPTI and persistence, and are evaluated based on their ability to forecast thunderstorms correctly. The primary performance metrics used to evaluate the algorithms are hit rate, threat score, probability of detection, false alarm rate, Brier skill score, and ratio skill score. The investigation results indicate that the new algorithms are suitable for use by the 45th Weather Squadron and are an improvement on NPTI. The results show that the best algorithm, Stratified Logistic Thunderstorm Index (STLI), has a 57% better hit rate, a 51% better threat score, and a 68% better probability of detection than NPTI. In Addition STLI shows a 59% lower false alarm rate than NPTI. Because of the significant improvement over NPTI, the algorithm should be prepared for operational use at Patrick AFB.

14. SUBJECT TERMS
meteorology, thunderstorms, instability, Florida, forecasting

15. NUMBER OF PAGES
103

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL